Wavelet Shrinkage in Signal & Image Processing

An Investigation of Relations and Equivalences

von Dirk A. Lorenz

Dissertation

zur Erlangung des Grades eines Doktors der Naturwissenschaften
– Dr. rer. nat. –

Vorgelegt im Fachbereich 3 (Mathematik & Informatik)
der Universität Bremen
im Oktober 2004

Datum des Promotionskolloquiums: 4. Februar 2005

Gutachter: Prof. Dr. Peter Maaß (Universität Bremen)
Prof. Dr. Otmar Scherzer (Universität Innsbruck)

Wavelet Shrinkage in Signal & Image Processing
An Investigation of Relations and Equivalences

von Dirk A. Lorenz

Zusammenfassung

Diese Arbeit ist ein Beitrag zum Themenfeld „Äquivalenzen von verschiedenen Methoden in der mathematischen Bildverarbeitung“. In den letzten zehn Jahren hat sich dieses Feld als eigenständiges Forschungsgebiet in der Mathematischen Bildverarbeitung etabliert. Die Arbeit präsentiert eine umfassende Untersuchung von Äquivalenz-Ergebnissen für spezielle Methoden des Entrauschens: die Wavelet-Shrinkage-Methoden.

Wavelet Methoden werden seit fast fünfzehn Jahren sehr erfolgreich in der Signal- und Bildverarbeitung angewendet und es hat sich in vielen Arbeiten gezeigt, dass Wavelet-Shrinkage-Methoden sich in verschiedenen Modellen zum Entrauschen „natürlich“ ergeben. Diese Ergebnisse kommen aus den unterschiedlichsten Gebieten: Harmonische Analysis, Funktionalanalysis, partielle Differentialgleichungen oder auch Statistik. Ziel dieser Arbeit ist es, all diese Ergebnisse in einem gemeinsamen Kontext zu präsentieren.

Neben diesen „klassischen“ Ergebnissen werden einige Verallgemeinerungen gezeigt, zum Beispiel: Hard und Soft Wavelet-Shrinkage lassen sich im selben Kontext behandeln und es ist sogar möglich, eine natürliche Interpolation zwischen beiden zu konstruieren; das abstrakte Konzept „Shrinkage“ kann auch auf andere Methoden zum Entrauschen angewendet werden wie zum Beispiel BV-Methoden oder Regularisierungen in Sobolev- oder Hölder-Räumen.

Abstract

This thesis is a contribution to the field “equivalences of different methods of mathematical image processing”. During the last decade this field has become an independent field of mathematical image processing. The intention of this thesis is to present an extensive collection of equivalence results for special denoising methods: the wavelet shrinkage methods.

Wavelet methods are applied in signal and image processing very successfully for almost fifteen years and it has been shown in several papers that wavelet shrinkage methods arise “naturally” in many different mathematical models for signal and image denoising. These results come from very different fields of mathematics: harmonic analysis, functional analysis, partial differential equations, or statistics. The aim of this thesis is to present all these equivalence results in a unifying framework.

Besides these “classical” results some generalizations are presented, for example: Hard and soft wavelet shrinkage can be treated in a common framework and it is possible to construct a natural interpolation between both of them; the abstract concept of “shrinkage” also applies to other methods for denoising, for example for BV denoising or even for regularizations which involve Sobolev or Hölder spaces.

I would like to thank:

Prof. Peter Maaß for support, motivation and for enforcing my ambition to write this thesis.

The whole “Zentrum für Technomathematik” for mathematical and technical support, especially the people of the “AG Technomathematik” for a good time and a pleasant environment, and even more special my colleagues Esther Klann, Henning Thielemann, Mathias Lindemann, Lutz Justen, Kristan Bredies and Gerd Teschke for advices, discussions and encouragement – both mathematically and personally.

Dirk A. Lorenz
Zentrum für Technomathematik
Universität Bremen

www.math.uni-bremen.de/~dlorenz

CONTENTS

CHAPTER 1

Introduction

The mathematical discipline of image processing makes use of many areas of mathematics: statistics, functional analysis, partial differential equations, calculus of variations, function-space theory, wavelets, or harmonic analysis are just a few. Any of these areas provides many tools for the main topics of what is called low level vision: denoising, deblurring, segmentation, and structure identification.

Lots of the methods from different areas are strongly connected and some are just the same in a different language. In the last years the investigation of the connections between different methods in image processing has become an own field of mathematical image processing.

This thesis is a contribution to this field of mathematical image processing. Its intention is to present an extensive collection of equivalence results for special denoising methods: the wavelet shrinkage methods.

Wavelet shrinkage is a quite recent denoising method compared to classical methods like the Wiener filter or convolution filters and is applied very successfully to various denoising problems. A very interesting thing about wavelet shrinkage is, that it can be motivated from very different fields of mathematics, namely partial differential equations, the calculus of variations, harmonic analysis or statistics. The aim of this thesis is to present all these motivations in one framework.

I would like to remark, that this thesis will present neither totally new methods nor optimized algorithms for denoising. Its aim is not to present a denoising method which outperforms every other method. In the spirit of mathematics it is about structure. The principles and the background

of the methods will be investigated and worked out. Special attention will be turned to similarities and differences of the underlying philosophies of different methods. Against this background one shall see the figures in this thesis as illustrations of the methods and compare them with the underlying motivations but not necessarily with the intention to find a best method.

1.1 Organization of the thesis

The thesis is divided into six chapters in which we address the multiple connections between different fields of image analysis and wavelet shrinkage methods.

In the introduction the wavelet transform and the notation which will be used throughout the thesis is briefly introduced. Further we give a first description and motivation of wavelet shrinkage, an outlook on the equivalence results which will be obtained and a short historical survey of shrinkage methods.

The Chapter 2 provides some theory which is needed and possibly not so widely known: convex analysis and the theory of Besov spaces. Besides the needed definitions and facts we provide motivations and illustrations to make this short introductions readable and understandable although the proofs are omitted. The reader familiar with these areas is invited to skip this chapter.

The connection between wavelet shrinkage and descent equations and other shrinkage methods is treated in Chapter 3. Starting point is the well known result of [CL01] which says, that discrete wavelet shrinkage is a descent along a certain semi norm. We provide more general results which cover a more general class of shrinkage methods and give more examples like continuous Fourier and wavelet shrinkage. Another result concerning wavelet shrinkage and differential equations is due to Weickert et al. [MWS03]. They discovered a connection between a special kind of wavelet shrinkage and some special diffusion equations in one and two dimensions. We present these results at the end of Chapter 3.

The Chapter 4 is divided in two sections. Both sections are concerned with the connection of wavelet shrinkage and variational methods. We show, that different variational functionals are closely related to different shrinkage methods. In the first section we obtain the result that soft and hard wavelet shrinkage can be treated in the same framework. In the second section it is shown how wavelet shrinkage methods are related to projection methods in Hilbert spaces.

In Chapter 5 the link between wavelet shrinkage and so called Bayesian methods from statistics is shortly discussed. The last chapter provides a summary of the obtained results.

1.2 What wavelet shrinkage is about

This section will give an introduction to the basic idea of wavelet shrinkage. The first subsections provide a short survey of the orthogonal wavelet transform on the real line, in higher dimensions, on cubes in $\mathbb{R}^d$, and of the continuous wavelet transform.

In the last subsection we provide an outline of the subject of wavelet shrinkage. We will give a first and very basic motivation and describe the history of shrinkage methods.

1.2.1 The orthogonal discrete wavelet transform

There are many different types of discrete wavelet transforms, but the basic one in the orthogonal wavelet transformation on the real line. We will start with this transformation and then we construct separable wavelets in $\mathbb{R}^d$. Finally we describe a type of wavelet expansion for functions $f \in L^2\left([0,1]^d\right)$ by periodization. More detailed introductions to the discrete wavelet transform can be found in [Dau92], [DL92b], [LMR97] or [Mal99].

Definition 1.1. *A function $\psi : \mathbb{R} \to \mathbb{R}$ is called an orthogonal wavelet, if the shifted and translated versions*

$$\psi_{j,k}(x) := 2^{j/2}\psi(2^j x - k)$$

form an orthonormal basis of $L^2(\mathbb{R})$.

The number j indicates the "scale" of the wavelet and k the position in space.

For an orthogonal wavelet ψ a function $f \in L^2(\mathbb{R})$ can be described through its wavelet coefficients

$$f_{j,k} := \langle f | \psi_{j,k} \rangle = \int_{\mathbb{R}} f(x)\psi_{j,k}(x)\mathrm{d}x.$$

Because $\{\psi_{j,k}\}$ is an orthonormal basis for $L^2(\mathbb{R})$ we have the wavelet expansion of f

$$f = \sum_{j,k\in\mathbb{Z}} f_{j,k}\psi_{j,k}$$

and the norm equivalence

$$\|f\|^2_{L^2(\mathbb{R})} = \sum_{j,k\in\mathbb{Z}} f^2_{j,k}.$$

In other words: The mapping

$$f \mapsto (f_{j,k})_{j,k\in\mathbb{Z}}$$

is an invertible isometry between the Hilbert spaces $L^2(\mathbb{R})$ and $\ell^2(\mathbb{Z}\times\mathbb{Z})$.

The most easy example for an orthogonal wavelet one can imagine is the so called Haar wavelet (named after Alfred Haar). The Haar wavelet is a piecewise constant function, defined by

$$\psi(x) = \begin{cases} 1 & ,0 \le x < 1/2 \\ -1 & ,1/2 \le x < 1 \\ 0 & , \text{ else.} \end{cases}$$

There are many examples for orthogonal wavelets (Meyer wavelets, Daubechies wavelets, Battle-Lemarié wavelets, so called coiflets (named after R. R. Coifman)...). For an overview the reader is referred to [Mal99].

Under additional assumptions on the wavelet there is a related scaling function ϕ (to be precise here, we had to introduce the concept of multiresolution analysis, for details on this we refer to [LMR97, Mal99]). The scaling function ϕ is orthonormal for a fixed scale, i. e. the set $\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ is orthonormal for fixed j. With the help of the scaling function we have another expansion for a function $f \in L^2(\mathbb{R})$. We fix a coarsest level j_0 and truncate the wavelet expansion. The part which is missing is captured by the scaling function on the level j_0:

$$f = \sum_{k\in\mathbb{Z}} \langle f | \phi_{j_0,k} \rangle \phi_{j_0,k} + \sum_{k\in\mathbb{Z}, j\ge j_0} f_{j,k}\psi_{j,k}.$$

The scaling function associated with the Haar wavelet is just the characteristic function on the unit interval:

$$\phi(x) = \chi_{[0,1[}(x).$$

We just mention that there a some generalizations of the wavelet transform, e. g. the biorthogonal wavelet transform, the wavelet packet transform or the stationary wavelet transform. Only the stationary wavelet transform will appear in this thesis. It is introduced in Section 3.3.

The coefficients $f_{j,k}$ are computed easily and fast with the help of filter banks (see [Mal99] for example). We use the following notations:

$$c^j(n) = \langle f | \phi_{j,n} \rangle , \; d^j(n) = \langle f | \psi_{j,n} \rangle$$

and call c^j the approximation coefficients of f at level j and d^j the detail coefficients of f at level j. Due to the scaling properties of the wavelets and the orthogonality of the wavelets and the scaling function one can show, that for orthogonal wavelets there exists a low pass filter $h(n)$ and a high pass filter $g(n)$ such that the approximation and the detail coefficients can be calculated with a cascade of discrete convolutions and subsamplings:

$$\begin{aligned} c^{j+1}(n) &= \sum_{k\in\mathbb{Z}} h(k-2n)c^j(k) = ((c^j * h) \downarrow 2)(n) \\ d^{j+1}(n) &= \sum_{k\in\mathbb{Z}} g(k-2n)c^j(k) = ((c^j * g) \downarrow 2)(n). \end{aligned}$$

The operator $*$ denotes in this case the discrete convolution and the operator $\downarrow 2$ is a downsampling by 2. Furthermore there is a simple reconstruction formula.

$$\begin{aligned} c^j(n) &= \textstyle\sum_{k\in\mathbb{Z}} h(n-2k)c^{j+1}(k) + \sum_{k\in\mathbb{Z}} g(n-2k)d^{j+1}(k) \\ &= ((c^{j+1} \uparrow 2) * \tilde{h})(n) + ((d^{j+1} \uparrow 2) * \tilde{g})(n). \end{aligned}$$

Where the filters $\tilde{h}$ and $\tilde{g}$ are low pass resp. high pass filters related to h and g and $\uparrow 2$ is an upsampling by 2 (for details, see [Mal99] for example).

Now we are going to describe separable wavelets in $\mathbb{R}^d$. This construction is probably the simplest way to get multivariate wavelets and is based on tensor products. We take all possible tensor products of the scaling function ϕ and the wavelet ψ. In $\mathbb{R}^d$ this procedure gives one scaling function and $2^d - 1$ wavelets. To do this construction precisely we need a little more notation.

We define $\phi_0 = \phi$ and $\phi_1 = \psi$ and consider the set $E = \{0,1\}^d$ which indicates the corners of a d-dimensional cube. For each corner $e \in E$ we set $\phi_e = \phi_{e_1} \otimes \cdots \otimes \phi_{e_d}$.

In two dimensions this looks like

$$\begin{aligned} \phi_{00}(x,y) &= \phi_0(x)\phi_0(y) = \phi(x)\phi(y) \\ \phi_{01}(x,y) &= \phi_0(x)\phi_1(y) = \phi(x)\psi(y) \\ \phi_{10}(x,y) &= \phi_1(x)\phi_0(y) = \psi(x)\phi(y) \\ \phi_{11}(x,y) &= \phi_1(x)\phi_1(y) = \psi(x)\psi(y) \end{aligned}$$

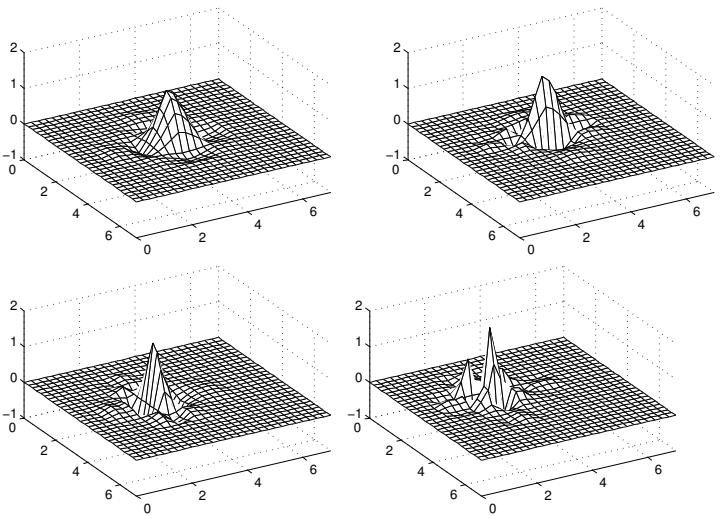

Figure 1.1: Two dimensional wavelets based on tensor products. The wavelet and scaling function are the so called symlet of order 4 wavelet resp. scaling function. The two dimensional scaling function is in the upper left corner, the other plots show the three wavelets.

and is illustrated in Figure 1.1

If we are working in $\mathbb{R}^d$ we use the following notation:

$$\begin{aligned} \phi &: \mathbb{R}^d \to \mathbb{R} \text{ scaling function} \\ \psi^i &\text{ wavelets } i = 1, \ldots, 2^d - 1. \end{aligned}$$

We define the dilations and translations for wavelets in $\mathbb{R}^d$ by

$$\psi^i_{j,k}(x) = 2^{jd/2}\psi^i(2^j x - k) \text{ for } k \in \mathbb{Z}^d, j \in \mathbb{Z}.$$

As before we make the abbreviation $f^i_{j,k} := \left\langle f \middle| \psi^i_{j,k} \right\rangle$. Then a function $f \in L^2\left(\mathbb{R}^d\right)$ has the following expansions

$$\begin{aligned} f &= \sum_{k\in\mathbb{Z}^d} \langle f | \phi_{0,k} \rangle \phi_{0,k} + \sum_{k\in\mathbb{Z}^d} \sum_{j=0}^{\infty} \sum_{i=1}^{2^d-1} f^i_{j,k}\psi^i_{j,k} \\ &= \sum_{k\in\mathbb{Z}^d} \sum_{j\in\mathbb{Z}} \sum_{i=1}^{2^d-1} f^i_{j,k}\psi^i_{j,k} \end{aligned}$$

where the first one is the wavelet series, truncated at level 0 and the second one is the full wavelet series. The discrete wavelet transform, which maps a function onto its wavelet coefficients, is an invertible isometry between the Hilbert spaces $L^2\left(\mathbb{R}^d\right)$ and $\ell^2\left(\mathbb{Z}^d \times \mathbb{Z} \times \{1, \ldots, 2^d - 1\}\right)$ if we think of the full wavelet series and between the Hilbert spaces $L^2\left(\mathbb{R}^d\right)$ and $\ell^2\left(\mathbb{Z}^d \cup \mathbb{Z}^d \times \mathbb{N} \times \{1, \ldots, 2^d - 1\}\right)$ if we think of a truncated expansion.

The filter bank algorithm in one dimension extends to the separable multivariate case (again, see [Mal99] for example).

Finally we describe wavelets on cubes in $\mathbb{R}^d$. There are different ways to define wavelets on cubes and we choose the way which uses periodization. This way can be found in [DL92b] or again in [Mal99].

The wavelet expansion of functions defined on $I := [0,1]^d$ is what we are most interested in, because one can think of an image as an element of $L^2(I)$. One way to deal with functions on I is to extend them on $\mathbb{R}^d$ by periodization. So it is natural to do so with the wavelets.

For a compactly supported wavelet (or scaling function) $\psi^i_{j,k}$ in $\mathbb{R}^d$ we define its periodic version by

$$\tilde{\psi}^i_{j,k}(x) = \sum_{l \in \mathbb{Z}^d} \psi^i_{j,k}(x - l).$$

One can show, that these periodic wavelets form an orthonormal basis of $L^2(I)$.

Remark 1.2. *For the scaling function we get*

$$\tilde{\phi} = \tilde{\psi}^0_{0,0} \equiv 1$$

because the translates of the scaling function form a partition of unity.

For the wavelet decomposition of functions in $L^2(I)$ we don't need all translates of these periodic wavelets. On the level j we only need the translations $k \in \Gamma_j := \{0, 1, \ldots, 2^j - 1\}^d$. The wavelet expansion for a function $f \in L^2(I)$ is

$$f = \left\langle f \middle| \tilde{\phi} \right\rangle + \sum_{j=0}^{\infty} \sum_{k \in \Gamma_j} \sum_{i=1}^{2^d - 1} \left\langle f \middle| \tilde{\psi}^i_{j,k} \right\rangle \tilde{\psi}^i_{j,k}.$$

Because of the above remark we write $\langle f | 1 \rangle$ for the scalar product $\left\langle f \middle| \tilde{\phi} \right\rangle$.

We are going to simplify the notation in the following. First we drop the "˜" if it is clear that we are working in I. To get rid of the triple sums it will be useful to collect the three indices in just one. We write $\gamma := (i, j, k)$ and

$\sum_{\gamma\in\Gamma}$ for the triple sum $\sum_{j=0}^{\infty}\sum_{k\in\Gamma_j}\sum_{i=1}^{2^d-1}$. Further we rewrite $\psi_\gamma = \psi_{j,k}^i$ for the dilated and translated wavelets and $f_\gamma := \langle f|\,\psi_\gamma\rangle$ for the wavelet coefficients.

With all these simplifications the wavelet expansion of $f \in L^2\left([0,1]^d\right)$ reads

$$f = \langle f|\,1\rangle + \sum_{\gamma\in\Gamma} f_\gamma\psi_\gamma.$$

1.2.2 The continuous wavelet transform

Like the Fourier transform the wavelet transform has a discrete and a continuous version. For the continuous wavelet transform one has weaker conditions for the wavelets, especially orthogonality in not necessary for an invertible continuous wavelet transform.

For the definition of an admissible wavelet we need the continuous Fourier transform $\mathcal{F}$. The reader unfamiliar with the Fourier transform can find the definition in Subsection 3.2.3 and a detailed introduction in [Rud87].

Definition 1.3 (Continuous wavelet transform). *Let $\psi \in L^2(\mathbb{R})$ be a function which fulfills the so called admissibility condition*

$$0 < c_\psi := 2\pi\int_{\mathbb{R}}\frac{|\mathcal{F}\psi(\omega)|^2}{|\omega|}d\omega.$$

The wavelet transform of $f \in L^2(\mathbb{R})$ with respect to ψ is

$$L_\psi f(a,b) = \frac{1}{\sqrt{c_\psi\,|a|}}\int_{\mathbb{R}} f(t)\psi\left(\frac{t-b}{a}\right)dt.$$

The continuous wavelet transform is an isometry between certain Hilbert spaces and provides an inversion formula (see [LMR97], for example).

Theorem 1.4 (Isometry and inversion formula). *The wavelet transform L_ψ maps $L^2(\mathbb{R})$ isometrically onto $L^2\left(\mathbb{R}^2, dadb/a^2\right)$. The adjoint of the wavelet transform*

$$L_\psi^*(g)(t) = \frac{1}{\sqrt{c_\psi}}\int_{\mathbb{R}}\int_{\mathbb{R}}\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)g(a,b)\frac{dadb}{a^2}$$

inverts the wavelet transform on its range.

1.2.3 Wavelet shrinkage - description and short history

A heuristic way to wavelet shrinkage goes as follows. We consider a signal f which is disturbed by additive white noise: $g = f + \varepsilon$. Since the discrete wavelet transform is linear and orthogonal, the wavelet transform of g has the form $(g_\gamma) = (f_\gamma) + (\varepsilon_\gamma)$ where the coefficients ε_γ of the noise are again white noise.

Usually the signal f results in a few number of large wavelet coefficients and most of the coefficients are zero or nearly zero. The noise on the other hand leads to a large number of small coefficients on all scales. Thus, the small coefficients in (g_γ) mostly contain noise.

Hence, it seems to be a good idea to set all the coefficients which are small to zero. But what shall happen to the large coefficients? There are a lot of different possibilities to answer this question. The two most popular ones are hard and soft shrinkage. By application of hard shrinkage one leaves the large coefficients unchanged and sets the coefficients below a certain threshold to zero. Mathematically spoken one applies the function

$$S_\lambda(x) = \begin{cases} x & , |x| > \lambda \\ 0 & , |x| \leq \lambda \end{cases}$$

to the wavelet coefficients. Another famous way is soft shrinkage where the magnitude of all coefficients is reduced by the threshold: One applies the function

$$S_\lambda(x) = \begin{cases} x - \lambda & , x \geq \lambda \\ 0 & , |x| \leq \lambda \\ x + \lambda & , x \leq -\lambda \end{cases}$$

to the coefficients.

Beside these two possibilities there are many others (semi-soft shrinkage, firm shrinkage,...) and as long as the shrinkage function preserves the sign ($\operatorname{sign}(S_\lambda(x)) = \operatorname{sign}(x)$) and shrinks the magnitude ($|S_\lambda(x)| \leq |x|$) one can expect a denoising effect.

The interesting thing about wavelet shrinkage is, that it appears in very different fields of mathematics in a natural way. It is the goal of this thesis to present four of the places where shrinkage appears naturally:

1. As the subgradient descent along the absolute value (Chapter 3).

2. As the function which maps an initial value onto the minimizer of a variational functional (in terms of variational analysis, this is the proximal mapping of the absolute value, see Section 4.1)

3. As the function "identity minus projection onto a convex set" which is also motivated by variational analysis (see Section 4.2).

4. As the maximum a posteriori estimator for an additively disturbed signal, where the signal and the noise are distributed in a certain way (see Chapter 5).

The first three items are worked out in detail and every item allows a few generalizations. The last item is presented briefly only for the sake of completeness.

To the best knowledge of the author, the first application of wavelet shrinkage methods in image and signal processing was the paper "Filtering Noise from Images with Wavelet Transforms" from 1991 [WXHC91] where a level dependent soft shrinkage is proposed as an edge preserving denoising method. The first thorough mathematical treatment of wavelet shrinkage was done by Donoho et al. in a series of technical reports in the early 1990s and published in [Don95, DJKP96, DJ98]. Donoho and his coworkers analyzed wavelet shrinkage methods in the context of minimax estimation and showed, that wavelet shrinkage generates asymptotically optimal estimates for noisy data that outperform any linear estimator. At the same time DeVore and Lucier studied wavelet shrinkage in terms of minimization problems with the help of K-functionals in [DL92a]. In Chapter 4 of this thesis we study wavelet shrinkage in this context. Some well known results and some generalizations are presented.

Further improvements of the understanding of wavelet shrinkage are due to the works [CDLL98, CL01] where wavelet shrinkage and translation invariant wavelet shrinkage are interpreted as smoothing scale spaces. In Chapter 3 we present a general framework which interprets various shrinkage methods as solutions of descent problems. This work is based on [BLM03].

Another way to understand and motivate wavelet shrinkage comes from statistics. The use of Bayesian estimators and wavelet transforms for image denoising together with assumptions on the distributions of the noise and the wavelet coefficients leads to shrinkage operations. An early work on this is [SA96] where shrinkage appeared under the name "wavelet coring".

CHAPTER 2

Basic theory

This chapter gives short introductions to two theories which are essential in this thesis: The theory which is usually called "convex analysis" and the theory of Besov spaces. To make the introductions more readable we omitted most of the proofs but added easy examples and illustrations.

2.1 Some convex analysis

In many parts of this thesis we make use of convex analysis. Some of the terms of convex analysis are not very widely used, so that we present the needed definitions and facts here.

This chapter is called "convex analysis", but this is probably not the best title. Some of the concepts, namely the subgradients and the notion of duality, hold for a more general class of functions. Rockafellar and Wets for example chose the term "variational analysis" for their book [RW98] on this topic. We use the term "convex analysis" nonetheless because it is popular, widely used and further convexity plays a very special role in variational analysis.

The facts presented in this chapter are classic in convex analysis. They can be found in [ET76] or in the more recent book [RW98] (but only in the finite dimensional case).

2.1.1 Basic definitions

We describe the notion of convexity and the related facts for real Banach spaces. This is a little more general than we will need it, (normally we will only need Hilbert spaces) but the notation is almost the same.

Let X be a real Banach space. A subset $C \subset X$ is called convex, if for $x, y \in C$ and $t \in [0,1]$ we have $tx + (1-t)y \in C$.

For convex mappings it is useful to use the completed real line.

Definition 2.1. *The completed real line is* $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. *We define an ordering on* $\overline{\mathbb{R}}$ *by the usual ordering of* $\mathbb{R}$ *and for* $r \in \mathbb{R}$ *we have* $-\infty < r < \infty$.

For $\emptyset \subset \overline{\mathbb{R}}$ *we define*

$$\sup \emptyset = -\infty \text{ and } \inf \emptyset = \infty.$$

A mapping $f : X \to \overline{\mathbb{R}}$ is called convex, if for $x, y \in X$ and $t \in [0,1]$ we have $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. For a convex mapping f on a Banach space X the domain is given by $\operatorname{dom} f = \{ x \in X | \, f(x) < \infty \}$. A convex function F is called proper, if $\operatorname{dom} f \neq \emptyset$.

A useful class of convex functions is given by the so called indicator functions:

Definition 2.2. *Let* $C \subset X$ *be a convex set. The function*

$$\delta_C(x) = \begin{cases} 0 & , x \in C \\ \infty & , x \notin C \end{cases}$$

is called indicator function of C.

Indicator functions can be used to describe constrained conditions or to include the domain of a convex function into the function itself. For example a minimization problem over a convex set can be described as a minimization problem over the whole space:

$$\operatorname*{argmin}_{x \in C} f(x) = \operatorname*{argmin}_{x \in X} f(x) + \delta_C(x).$$

The next term we need is the notion of lower semi continuity which is important for the existence of minimizers.

Definition 2.3. *A function* $f : X \to \overline{\mathbb{R}}$ *is lower semi continuous if for all* $\bar{x} \in X$

$$\liminf_{x \to \bar{x}} f(x) \left(:= \lim_{\varepsilon \to 0} \inf_{\|x - \bar{x}\|_X < \varepsilon} f(x) \right) = f(\bar{x}).$$

Remark 2.4. *An equivalent definition for lower semi continuity is:*

$$x_n \to x \text{ in } X \text{ and } f(x_n) \leq a \implies f(x) \leq a.$$

Example 2.5. *The function* $\delta_{[-1,1]} : \mathbb{R} \to \overline{\mathbb{R}}$ *is lower semi continuous but the function* $\delta_{]-1,1[}$ *is not. One easily sees, that* $\liminf_{x\to 1} \delta_{]-1,1[}(x) = 0 \neq \infty = \delta_{]-1,1[}(1)$*, for example.*

2.1.2 Set valued mappings

In convex analysis it comes quite naturally (as we will see in the next section about subgradients) that one has to consider set valued mappings, i. e. mappings where one point is mapped on a set of points.

The first definition that comes to mind for a set valued mapping is: A set valued mapping from a set A to another set B has the form $f : A \to \mathfrak{P}(B)$ where $\mathfrak{P}(B)$ denotes the power-set of B.

But this does not fit our intuition that a set valued mapping describes a "varying set depending on a variable" very well. This intuition can be modeled by the graph of a function: We think of the graph of a set valued mapping as a subset of $A \times B$ (and not as a subset of $A \times \mathfrak{P}(B)$ as it would be for the first definition):

$$\operatorname{graph} f = \{\, (a, b) \in A \times B |\, b \in f(a) \}\,.$$

The other way round, a set valued mapping is completely described by its graph $G \subset A \times B$:

$$f(a) = \{\, b \in B |\, (a, b) \in G \}$$

and this is the way we think of set valued mappings (see Figure 2.1 for illustration).

Some authors use the notation $f : A \rightrightarrows B$ for set valued mappings. In this thesis we will not do this and use the same notation for set valued mappings as for mappings and hopefully this will not lead to confusions.

Definition 2.6 (Domain and inverse of a set valued mapping). *The domain of a set valued mapping* f *is*

$$\operatorname{dom} f = \{\, a \in A |\, f(a) \neq \emptyset \}$$

The inverse of a set valued mapping is

$$f^{-1}(b) = \{\, a \in A |\, b \in f(a) \}\,.$$

In other areas of mathematics these objects are called "relations" but this is a point of view which is not adapted to what we want of set valued mappings.

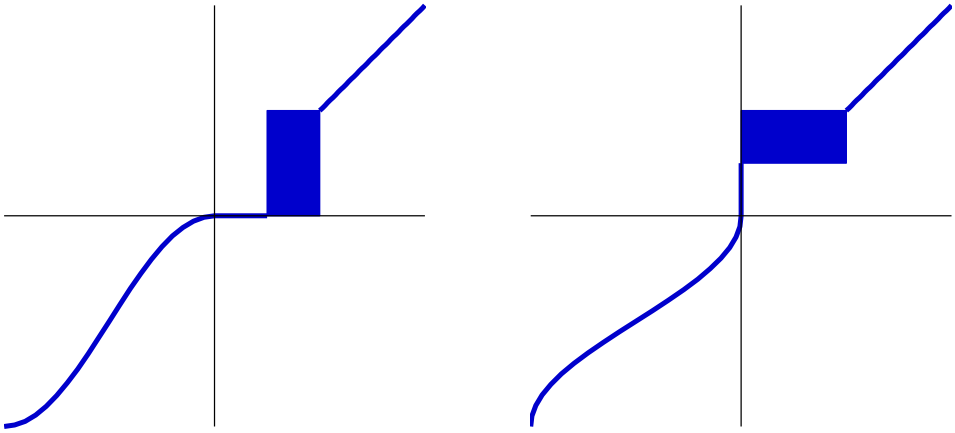

Figure 2.1: A set valued mapping $f : \mathbb{R} \to \mathbb{R}$ and its inverse.

2.1.3 Subgradients and duality

This section describes important objects and powerful concepts from convex analysis we need in the following chapters: subgradients and the notion of duality.

Subgradients are a generalization of the usual gradient or derivative. If one thinks of the gradient as a vector which describes a supporting hyperplane, the subgradient is the set of all vectors which do the same. So the subgradient is in general a set valued mapping.

One of the most fruitful notions in convex analysis is the notion of duality. There are several concepts of duality. We are going to describe only two types of duality: The Fenchel duality of functions and the duality between convex sets and positive homogeneous functionals. The second one is actually a special case of the first one.

We begin with the definition of the subgradient.

Definition 2.7 (Subgradient). *Let $f : X \to \overline{\mathbb{R}}$ be a mapping from a Banach space X into the extended real line. The subgradient $\partial f(\bar{x})$ of f at $\bar{x}$ is a subset of the dual space X^* defined by*

$$x^* \in \partial f(\bar{x}) \iff \liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \langle x^* | \, x - \bar{x} \rangle_{X^* \times X}}{\|x - \bar{x}\|_X} \geq 0.$$

The domain of the subgradient is

$$\operatorname{dom} \partial f = \{ x \in X | \, \partial f(x) \neq \emptyset \}.$$

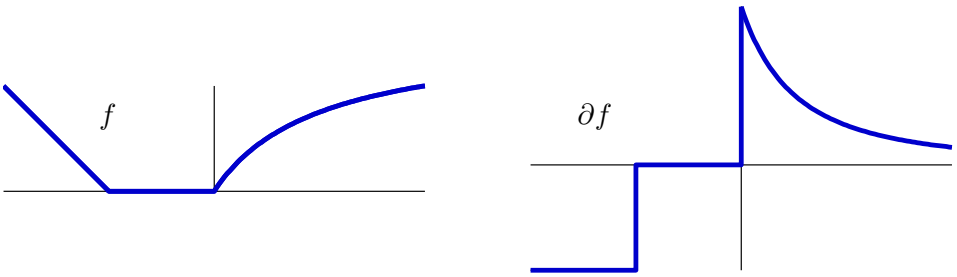

Figure 2.2: A function $f : \mathbb{R} \to \mathbb{R}$ and its subgradient.

For convex functions the definition is a little more easy:

Proposition 2.8 (The subgradient of a convex funtion). *Let $f : X \to \overline{\mathbb{R}}$ be a convex function. Then x^* is in the subgradient of f at $\bar{x}$ if and only if*

$$f(x) \geq f(\bar{x}) + \langle x^* | \, x - \bar{x} \rangle_{X^* \times X}, \forall x \in X.$$

If a function is differentiable (in the sense of Gâteaux) then its subgradient is single valued and coincides with the Gâteaux derivative. So one has $\partial f(x) = \{f'(x)\}$. Figure 2.2 illustrates this fact.

Example 2.9 (Subgradient of the absolute value). *We give a simple example which we will need in Chapter 3. Let $X = \mathbb{C}$ and $f(z) = |z|$. Because we developed the theory for real Banach spaces we make the identification $\mathbb{C} \simeq \mathbb{R}^2$. This gives a different scalar product in $\mathbb{C}$: $\langle x + iy | \, u + iv \rangle = xu + yv$.*

As one can check the subgradient of f is

$$\partial f(z) = \begin{cases} \frac{z}{|z|} & \text{for } z \neq 0 \\ \{\zeta \in \mathbb{C} | \, |\zeta| \leq 1\} & \text{for } z = 0. \end{cases}$$

In Section 3.1 we will see more difficult examples in infinite dimensional Banach spaces.

Now we define the Fenchel transformation (which is often called Fenchel-Legendre transformation and in special cases only Legendre transformation).

Definition 2.10 (Fenchel transformation). *Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$ a functional. The Fenchel transform (or dual function) of f is $f^* : X^* \to \overline{\mathbb{R}}$ and is defined by*

$$f^*(x^*) = \sup_{x \in X} \left(\langle x^* | \, x \rangle_{X^* \times X} - f(x) \right).$$

Remark 2.11. *Usually this definition is only given for convex, proper, lower semi continuous functionals. But it causes no problems to state it for all functionals. The convex, proper and lower semi continuous functionals play a special role concerning the Fenchel transformation. It holds:*

$$f \text{ convex, proper and lower semi continuous } \implies f^* \text{ as well.}$$

and furthermore

$$f \text{ convex, proper and lower semi continuous } \Leftrightarrow f = f^{**}.$$

Example 2.12 (Dual functions of indicator functions). *Let $C \subset X$ be a convex set. Then the dual of the indicator function δ_C of C is*

$$(\delta_C)^*(x^*) = \sup_{x \in C} \langle x^* | \, x \rangle \, .$$

The function $(\delta_C)^*$ clearly is a positive homogeneous function.

One can show, that for every positive homogeneous function $f : X \to \overline{\mathbb{R}}$ there exists a set $C^* \subset X^*$ such that

$$f = (\delta_{C^*})^*.$$

The set C^* has the form

$$C^* = \left\{ x^* \in X^* \middle| \, \langle x^* | \, x \rangle_{X^* \times X} \leq f(x) \ \forall x \right\}.$$

If X is a Hilbert space which is identified with its dual then there is a total duality between convex sets and positive homogeneous functionals: For every convex set $C \in X$ there is a corresponding positive homogeneous functional $f = (\delta_C)^*$ which is called the "support functional" and for every positive homogeneous functional f there is a corresponding convex set C described as above.

Proof. We will shortly sketch the proof of the stated assertions. At first it is clear, that the dual function of a one homogeneous functional only takes the values 0 and ∞, because it holds $\lambda f^* = f^*$ for every $\lambda > 0$.

If we assume $f^*(x^*) = 0$ it follows, that $\sup_{x \in X} \langle x^* | \, x \rangle_{X^* \times X} - f(x) = 0$ and hence $\langle x^* | \, x \rangle_{X^* \times X} \leq f(x)$. This shows

$$f^*(x^*) = 0 \implies x^* \in C^*.$$

For the other way round we assume $x^* \in C^*$ which yields $f^*(x^*) = \sup_{x \in X} \langle x^* | \, x \rangle_{X^* \times X} - f(x) \leq 0$ and since f^* only takes the values 0 and ∞ is has to be 0. □

The following lemma characterizes the subgradient in terms of the dual function.

Lemma 2.13 (Characterization of the subgradient). *Let $f : X \to \overline{\mathbb{R}}$ be proper and convex. Then it holds*

$$\bar{x}^* \in \partial f(\bar{x}) \iff f(\bar{x}) + f^*(\bar{x}^*) = \langle \bar{x}^* | \bar{x} \rangle_{X^* \times X} .$$

Proof. First we mention that the inequality

$$f(x) + f^*(x^*) \geq \langle x^* | x \rangle_{X^* \times X}$$

holds for all x, x^*.

To show the opposite inequality we observe that $\bar{x}^* \in \partial f(\bar{x})$ implies

$$f(x) \geq f(\bar{x}) + \langle \bar{x}^* | x - \bar{x} \rangle_{X^* \times X} \text{ for all } x.$$

This is equivalent to

$$\begin{aligned} & \langle \bar{x}^* | \bar{x} \rangle_{X^* \times X} - f(\bar{x}) \geq \langle \bar{x}^* | x \rangle_{X^* \times X} - f(x) \text{ for all } x \\ \Leftrightarrow\ & \langle \bar{x}^* | \bar{x} \rangle_{X^* \times X} - f(\bar{x}) \geq \sup_x (\langle \bar{x}^* | x \rangle_{X^* \times X} - f(x)) \\ \Leftrightarrow\ & \langle \bar{x}^* | \bar{x} \rangle_{X^* \times X} - f(\bar{x}) \geq f^*(\bar{x}^*) \end{aligned}$$

which shows the inequality. □

The last fact we present in this chapter about convex analysis follows directly from the above lemma. The theorem says how the inverse of the subgradient is expressed in terms of the dual function.

Theorem 2.14 (Inversion rule for subgradients). *Let X be a Banach space and $f : X \to \overline{\mathbb{R}}$ a proper, convex and lower semi continuous functional. Then*

$$(\partial f)^{-1} = \partial (f^*) .$$

Proof. By the definition of the inverse of a set valued mapping and by Lemma 2.13 we have the following sequence of equivalences

$$\begin{aligned} x \in (\partial f)^{-1}(x^*) \quad & \Leftrightarrow \quad x^* \in \partial f(x) \Leftrightarrow f(x) + f^*(x^*) = \langle x^* | x \rangle_{X^* \times X} \\ & \Leftrightarrow \quad x \in \partial (f^*)(x^*). \end{aligned}$$

□

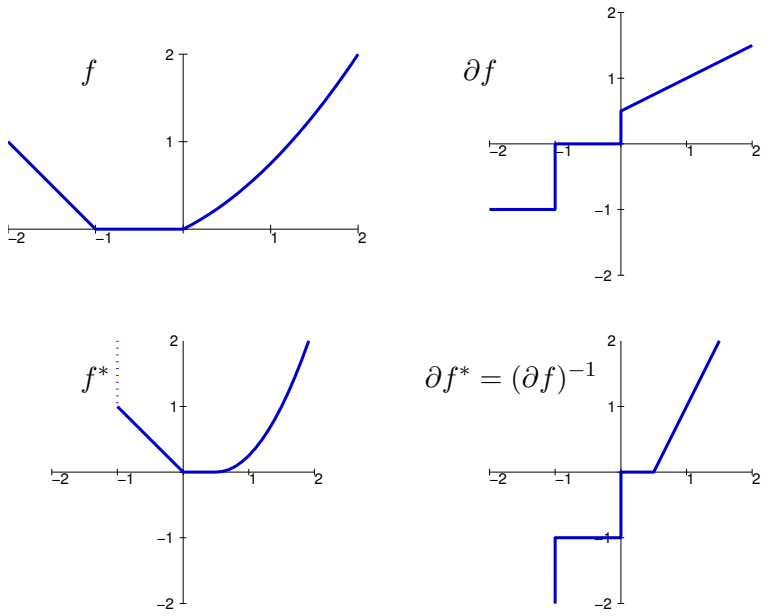

Figure 2.3: Illustration of the inversion rule for subgradients.

Example 2.15 (The absolute value again). *An very simple example for the inversion rule and also for the duality between positive homogeneous functionals and convex sets is the absolute value.*

We consider the functional $f(x) = |x|$*. The subgradient is given by the sign:*

$$\partial f(x) = \operatorname{sign}(x) = \begin{cases} -1 & , \ x < 0 \\ [-1,1] & , \ x = 0 \\ 1 & , \ x > 0 \end{cases}.$$

As one can check easily the dual function of f *is* $f^*(x) = \delta_{[-1,1]}(x)$*. Then the subgradient of* f^* *can either be computed directly from* f^* *or simply as the inverse function of the sign which is*

$$\partial(f^*)(x) = (\operatorname{sign})^{-1}(x) = \begin{cases} \emptyset & , \ |x| > 1 \\ 0 & , \ |x| < 1 \\ ]-\infty, 0] & , \ x = -1 \\ [0, \infty[& , \ x = 1 \end{cases}.$$

Another graphical example for the inversion rule is shown in Figure 2.3.

2.2 Introduction to Besov spaces

This chapter gives a very short introduction to Besov spaces. Besov spaces are function spaces introduced by Besov in the 1950s which measure smoothness and integrability in a very wide sense. Especially they cover some common smoothness spaces like Hölder spaces and Sobolev spaces. An excellent summary of the development and history of Besov and related function spaces is [Tri92].

The special role of Besov spaces for wavelet shrinkage comes from the fact that the Besov norms which measure smoothness have equivalent descriptions by means of wavelet coefficients.

There are different ways to introduce Besov spaces. Here we have chosen the definition via moduli of smoothness. Further, we describe the construction of Besov spaces only on special domains, namely the domains $I := [0,1]^d \subset \mathbb{R}^d$.

2.2.1 Moduli of smoothness

A very basic idea to measure the smoothness of a function is to consider differences of the function and a translated version of the function.

Definition 2.16 (Iterated discrete differences). *Let $f : I \to \mathbb{R}$, $h \in \mathbb{R}^d$. We define the difference operators by*

$$\Delta_h^0 f(x) := f(x)$$
$$\Delta_h^{k+1} f(x) := \Delta_h^k f(x+h) - \Delta_h^k f(x).$$

For $r > 0$, $\Delta_h^r f$ is only defined for

$$x \in I_{rh} := \{ x \in I | \, x + rh \in I \}.$$

Remark 2.17. *By the binomial theorem one gets the closed form expression for the r-th discrete difference:*

$$\Delta_h^r f(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f(x+kh).$$

The behavior of the discrete differences for small h describes smoothness of a function. The faster the discrete difference "approaches zero", the smoother the function is. The way how the difference "approaches zero" can be measured in different norms. This information is covered by the so called modulus of smoothness which uses the Lebesgue norms.

Definition 2.18 (Modulus of smoothness). *Let $0 < p \leq \infty$. The $L^p(I)$-modulus of smoothness of order r is defined as*

$$\omega_r(f,t)_p := \sup_{|h|\leq t} \left(\int_{I_{rh}} |\Delta_h^r f(x)|^p \, dx \right)^{1/p}.$$

For $p = \infty$ one has to replace the integral by a supremum as usual.

Example 2.19 (Moduli of smoothness of one-sided monomials). *We consider the function $f : [-1,1] \to \mathbb{R}$ given by*

$$f(x) = x^n H(x) = \begin{cases} 0 & , x < 0 \\ x^n & , x \geq 0 \end{cases}$$

which we call one-sided monomial of degree $n \in \mathbb{N}$. H denotes the Heaviside function. These one-sided monomials are prototypes for functions which n-th derivative is not continuous. Here we are going to calculate their moduli of smoothness. In the next subsection we will use this result to figure out in what Besov space these one-sided monomials are contained in.

First we consider the function f for $n = 0$, i. e. the Heaviside function itself. The discrete differences $\Delta_h^k f$ are defined on $[-1, 1-kh]$ and are given by

$$\Delta_h^k f(x) = \begin{cases} 0 & , x < -kh \\ \sum_{l=j}^{k} (-1)^{k-l} \binom{k}{l} & , -jh \leq x < -(j-1)h \\ \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} & , x \geq 0. \end{cases}$$

The sum is 0 for $x \geq 0$ because it is an alternating sum of binomial coefficients. Further $\Delta_h^k f$ is constant on the intervals $]-jh, -(j-1)h[$ and the constant depends only on j and k. We denote this constant by $C(j,k)$.

To calculate the modulus of smoothness of f we investigate the integral:

$$\begin{aligned} \int_{-1}^{1-kh} \left| \Delta_h^k f(x) \right|^p dx &= \sum_{j=0}^{k} \int_{-jh}^{-(j-1)h} |C(j,k)|^p \, dx \\ &= h \sum_{j=0}^{k} |C(j,k)|^p . \end{aligned}$$

Thus, the modulus of smoothness is

$$\omega_k(f,t)_p = C(k,p) t^{1/p},$$

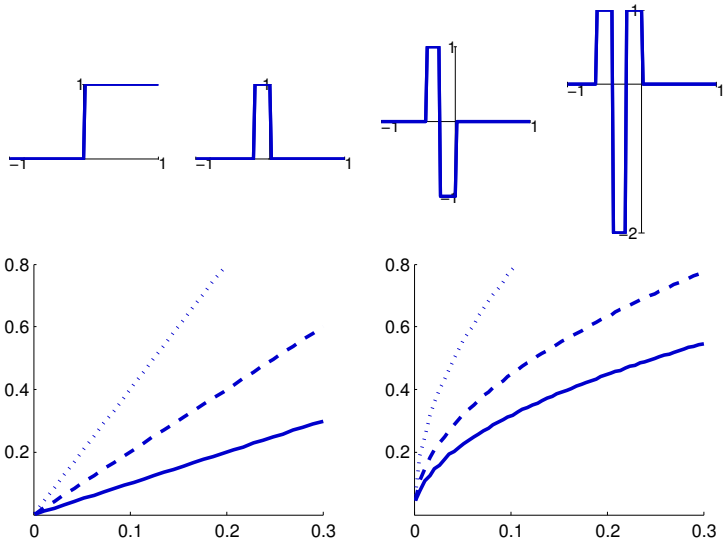

Figure 2.4: Upper row: The zeroth to third discrete difference of the one-sided monomial of degree zero (the Heaviside function) for $h = 0.2$. Lower row: The modulus of smoothness of the Heaviside function H. Left: $\omega_n(H,t)_1$, right: $\omega_n(H,t)_2$. The solid lines stand for first differences ($n = 1$), the dashed lines for $n = 2$ and the dotted lines for $n = 3$. Notice that the asymptotic behavior for $t \to 0$ does not depend on n but on p.

with another constant $C(k,p)$ which depends only on k and p.

To figure out how the moduli of smoothness of the higher order one-sided monomials look like, we use the following estimate

$$\omega_{r+k}(f,t)_p \leq t^r \omega_k(f^{(r)},t)_p$$

for the modulus of smoothness of the derivative(see [DL93] for this and more general results). If we consider $f(x) = x^n H(x)$ then we have $f^{(n)}(x) = n!H(x)$. It follows that the modulus of smoothness of f satisfies the inequality

$$\omega_{n+k}(f,t)_p \leq t^n \omega_k(n!H,t)_p = C(k,p,n)t^{n+1/p}.$$

Figure 2.4 shows an illustration of the discrete differences and the modulus of smoothness.

2.2.2 Norms and Besov spaces

Now we come to the definition of the Besov spaces. It involves the moduli of smoothness for a dyadic set of numbers. The definition is as follows.

Definition 2.20 (Besov spaces). *Let $s > 0$, $0 < p, q \leq \infty$ and $n \in \mathbb{N}$ such that $n > s$. A function $f : I \to \mathbb{R}$ belongs to the Besov space $B^s_{p,q}(I)$ if $f \in L^p(I)$ and the Besov semi norm, defined by*

$$|f|_{B^s_{p,q}(I)} := \left\| \left(2^{sj} \omega_n(f, 2^{-j})_p\right)_{j \geq 0} \right\|_{\ell^q(\mathbb{N})}$$

is finite. The Besov norm is

$$\|f\|_{B^s_{p,q}(I)} = \|f\|_{L^p(I)} + |f|_{B^s_{p,q}(I)}.$$

Remark 2.21. *The number n which appears in the definition of the Besov spaces is not important for the definition (see Figure 2.4). Different values of n give equivalent norms. One could take n such that $n - 1 \leq s < n$.*

The Besov spaces are Banach spaces for $p, q \geq 1$. For $p, q < 1$ the norm defined as above is only a quasi norm, i. e. the triangle inequality is not satisfied but it holds

$$\|f + g\| \leq C(\|f\| + \|g\|)$$

for some C independently of f and g.

Details on this remarks can be found in [Coh03] for example.

Example 2.22 (In which Besov space are the one-sided monomials?). *Again we consider the one-sided monomials. Let f be the one-sided monomial of degree n*

Example 2.19 in the previous subsection shows that the modulus of smoothness of f is estimated by

$$\omega_r(f, t)_p \leq C(r, p, n) t^{n+1/p}$$

if $r > n$.

Now we want to find out in what Besov space these one-sided monomials are contained. To do so we take a look at the Besov semi norm and out when it is finite.

$$\begin{aligned} |f|^q_{B^s_{p,q}([-1,1])} &= \left\| \left(2^{sj} \omega_{n+1}(f, 2^{-j})_p\right)_{j \in \mathbb{N}} \right\|^q_{\ell^q(\mathbb{N})} \\ &\leq \sum_{j \in \mathbb{N}} \left(2^{sj} C(r, p, n) 2^{-j(n+1/p)}\right)^q \\ &= C(r, p, n)^q \sum_{j \in \mathbb{N}} 2^{jq(s-n-1/p)}. \end{aligned}$$

The endmost sum is finite if and only if $s < n+1/p$ and $q < \infty$. In the case $q = \infty$ the condition on s and p is $s \leq n + 1/p$.

As a result we get that every degree of classical differentiability gives one degree of Besov smoothness.

Another result we want to mention is, that one dimensional functions with jump discontinuities belong to the Besov space $B^1_{1,\infty}(\mathbb{R})$ but not to $B^1_{1,q}(\mathbb{R})$ for every $q < \infty$.

2.2.3 Special cases and dual spaces

The dual spaces of Besov spaces for $1 \leq p, q \leq \infty$ are denoted by

$$\left(B^s_{p,q}(I)\right)^* = B^{-s}_{p^*,q^*}(I)$$

where p^* and q^* are the dual exponents of p and q and are defined through

$$\frac{1}{p} + \frac{1}{p^*} = 1, \quad \frac{1}{q} + \frac{1}{q^*} = 1.$$

As mentioned at the beginning of this chapter, some Besov spaces are just known function spaces (see [Mey92, Tri92], for example).

Proposition 2.23 (Hölder spaces are Besov spaces). *For $s > 0$ the spaces $B^s_{\infty,\infty}(I)$ are the known Hölder spaces $C^s(I)$.*

Proposition 2.24 (Sobolev spaces are Besov spaces). *For $s \in \mathbb{R}$ the Besov space $B^s_{2,2}(I)$ is equivalent to the Sobolev space $W^{s,2}(I) = H^s(I)$.*

2.2.4 Equivalent wavelet norms

Besov spaces are very useful in image processing because of two facts:

1. They provide a precise way to define smoothness. Moreover they capture a lot of classical smoothness classes as we have seen above. This makes Besov spaces an appropriate tool for modeling smoothness, oscillations and other features which are of interest in image processing.

2. There is an easy description for the Besov spaces and their norms through the coefficients of wavelet expansions. This simplifies theory and computations and especially makes computations fast.

The second point is essential in this thesis. There are many different results on various characterizations of Besov spaces by means of wavelet coefficients (for example [DL92b, DL93, RS96, DeV98, DKW98, Coh03]). We only quote the following one from [FJW91] which involves periodic wavelets as constructed in subsection 1.2.1.

Theorem 2.25 (Description of Besov spaces by means of wavelets). *Let $s > 0$ and $1 \leq p, q \leq \infty$ and let ψ be the Meyer wavelet (for the exact conditions see [FJW91]). Then the Besov semi norm as introduced in Definition 2.20 has the following equivalent description:*

$$|f|_{B^s_{p,q}(I)} \asymp \left(\sum_j \left(2^{sjp} 2^{j(p-2)d/2} \sum_{i,k} |f_\gamma|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}$$

where $f_\gamma = \langle f | \psi_\gamma \rangle$ are the wavelet coefficients of f corresponding to the periodic wavelet expansion.

This theorem makes Besov spaces and norms applicable in computations and in the following we will only use the Besov norms in terms of wavelet coefficients.

To the best knowledge of the author there is no reference where one can find an equivalence result for periodic wavelets with weaker conditions on the wavelet or for all other cases s, p and q. Nevertheless it seems to be true that the Besov spaces $B^s_{p,q}(I)$ are characterized through sequence norms of wavelet coefficients by the same formula provided in the above theorem for $s \in \mathbb{R}$ and $0 < p, q \leq \infty$ if the wavelet has a high regularity say it is in $B^r_{p,q}(I)$ for $r > s$. In the following we will use such wavelet norms for all Besov spaces which will appear even if it not clear if they describe the certain Besov space. One could think of the corresponding spaces as "function spaces close to Besov spaces".

Special cases of this theorem which will be of interest are:

$$|f|_{C^s(I)} \asymp |f|_{B^s_{\infty,\infty}(I)} \asymp \sup_{\gamma \in \Gamma} \left(2^{j(s+d/2)} |f_\gamma| \right)$$

$$|f|_{H^s(I)} \asymp |f|_{B^s_{2,2}(I)} \asymp \left(\sum_{\gamma \in \Gamma} 2^{2sj} |f_\gamma|^2 \right)^{\frac{1}{2}}$$

$$|f|_{B^s_{1,1}(I)} \asymp \sum_{\gamma \in \Gamma} 2^{j(s-d/2)} |f_\gamma| \,.$$

CHAPTER 3

Wavelet shrinkage and descent equations

The connection between soft wavelet shrinkage and descent equations was pointed out in [CL01]. In this chapter we show that there is a general theory for soft shrinkage methods and descent. The first main result is that a shrinkage of the values of a function (rather than the wavelet coefficients) is exactly the descent along an L^1-Norm. Then we show how this can be employed for shrinkage after applying an isometrical transformation as it is the case for the classical wavelet shrinkage.

We investigate the cases where the isometrical transformation is either the discrete wavelet transform, the discrete or the continuous Fourier transform. Problems arise if the isometrical transformation is not onto or if the range of the transform is not invariant under shrinkage (a weaker condition).

This is the case for the continuous wavelet transform and the stationary wavelet transform, which which are both redundant transformations. For the continuous wavelet transform we work out in detail what is going wrong if one tries to interpret the shrinkage as a descent.

The stationary wavelet transform allows different solutions. One possibility is presented in [CL01] and we refer to this paper. Another one is totally different and has been obtained by Mrazek et al. in [MWS03]. Because their approach allows another different interpretation of wavelet shrinkage, we present the results briefly in the last section of this chapter. It is shown, that shrinkage of the stationary wavelet transform is related to diffusion equations with certain diffusivities.

Most of the theory in the first part of the chapter can be found in [BLM03] and early results are in [Lor02].

3.1 Descent along L^1 norms

The relation between shrinkage and descent equations is based on a very simple fact. This basic observation, which motivates this chapter, is the following: Consider the ordinary differential equation (which is actually a differential inclusion)

$$\dot{x}(t) \in -\operatorname{sign}(x(t)), \qquad x(0) = x_0.$$

It is an inclusion, because the operator sign is meant to be the subgradient of the absolute value, i. e. $\operatorname{sign}(x) = \partial |x|$ (see Example 2.9). As one can check easily, this equation has a unique solution which is given by

$$x(t) = \begin{cases} x_0 - t & \text{for } t < x_0 \\ 0 & \text{for } t \geq x_0 \end{cases}$$

for an initial value $x_0 > 0$ (for $x_0 < 0$ the situation is converse). Some of these solutions are plotted in Figure 3.1 together with the directional field of the differential inclusion.

The so-called "evolution operator" S_t which maps the initial value x_0 on the value of the corresponding solution at time t is

$$S_t(x_0) = \begin{cases} x_0 - t & \text{for } x_0 > t \\ 0 & \text{for } -t \leq x_0 \leq t \\ x_0 + t & \text{for } x_0 < -t \end{cases}$$

and this is the soft shrinkage function.

Thus, the evolution of this differential inclusion describes a soft shrinkage of the initial value. Or, in other words: *The soft shrinkage is the subgradient descent along the absolute value.* This is the first place where we meet the soft shrinkage function in a context completely different from denoising of images and signals.

From this point of view it is natural to ask if wavelet shrinkage can be seen as the solution of a differential inclusion where the initial value is the given signal or image and the right hand side is the negative subgradient of a certain functional.

The first sections give a positive answer to this question. The first step is to generalize the above observation to infinite dimensional Hilbert spaces.

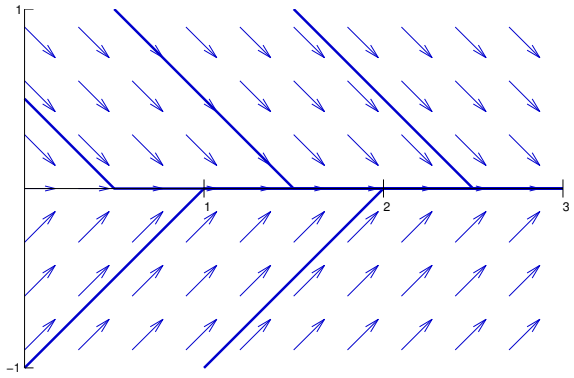

Figure 3.1: Some solutions of the differential inclusion $\dot{x} + \partial |x| \ni 0$ for different initial values.

3.1.1 The L^1 norm in L^2 and its subgradient

As we will see in the following, the right generalization of the above observation involves L^1 norms. We assume that $(\Omega, \mathrm{d}\mu)$ is a σ-finite measure space. The space $L^p(\Omega, \mathbb{C}, \mathrm{d}\mu)$ is the well known Lebesgue space with the canonical norm.

In the case $p = 2$ we use the following inner product which is a little unusual: We identify the space $L^p(\Omega, \mathbb{C}, \mathrm{d}\mu)$ with the real space $[L^p(\Omega, \mathbb{R}, \mathrm{d}\mu)]^2$. This gives the same norm as above, but a different scalar product:

$$\langle v | w \rangle_{L^p(\Omega,\mathbb{C},\mathrm{d}\mu)} = \int_\Omega \langle v(x) | w(x) \rangle_{\mathbb{R}^2} \, \mathrm{d}\mu$$

(compare Example 2.9).

In the following we only write $L^p(\Omega, \mathrm{d}\mu)$ for either $L^p(\Omega, \mathbb{C}, \mathrm{d}\mu)$ or $L^p(\Omega, \mathbb{R}, \mathrm{d}\mu)$ because the results hold for both cases. If it is necessary, we will state the differences explicitly.

On the space $L^2(\Omega, \mathrm{d}\mu)$ we define the functional $\Phi : L^2(\Omega, \mathrm{d}\mu) \to \overline{\mathbb{R}}$ by

$$\Phi(v) = \begin{cases} \int_\Omega |v| \, \mathrm{d}\mu & \text{where the integral exists and is finite} \\ \infty & \text{else.} \end{cases}$$

We will often use the more suggestive notation $\Phi(v) = \|v\|_{L^1(\Omega,\mathrm{d}\mu)}$.

We are going to show that pointwise shrinkage of an initial value $v \in L^2(\Omega, \mathrm{d}\mu)$ is a subgradient descent along Φ, i. e. the $L^1(\Omega, \mathrm{d}\mu)$ norm is the right generalization of the absolute value in $\mathbb{R}$.

Lemma 3.1 (Properties of the L^1 norm in L^2). *The functional Φ defined as above is proper, convex and lower semi continuous.*

Proof. We have $\Phi(0) = 0$ so that Φ is a proper functional. Φ is a norm on $\operatorname{dom}\Phi$ and hence it is convex.

The lower semi continuity of Φ is also classical and can be shown with the definition given by Remark 2.4 and the help of Fatou's lemma. □

To calculate the solution of the descent along the L^1 norm we need to know the subgradient of the L^1 norm explicitly. We give the following lemma which tells us how the subgradient of a class of integral functionals looks like. The subgradient of the L^1 norm is just a special case and will be given in the next corollary.

Lemma 3.2 (Subgradients of integral functionals). *Let $f : \mathbb{C} \to \mathbb{R}$ be convex and let $F : L^2(\Omega, d\mu) \to \overline{\mathbb{R}}$ be defined by*

$$F(u) = \begin{cases} \int_\Omega f(u)d\mu & \text{where the integral exists and is finite} \\ \infty & \text{else.} \end{cases}$$

Then $v \in L^2(\Omega, d\mu)$ is an element of $\partial F(u)$ if and only if $v(x) \in \partial f(u(x))$ for almost every $x \in \Omega$.

Proof. We calculate the subgradient with the help of the characterization through the dual function (Lemma 2.13). In this special case the dual function is given by

$$F^*(v) = \int_\Omega f^*(v(x))\mathrm{d}\mu$$

(see for example [ET76]).

The characterization of the subgradient in terms of the dual function says that

$$v \in \partial F(u) \Leftrightarrow F(u) + F^*(v) = \langle u | v \rangle_{L^2(\Omega,\mathrm{d}\mu)}.$$

This implies $v \in \partial F(u)$ if and only if

$$\int_\Omega f(u(x)) + f^*(v(x)) - \langle u(x) | v(x) \rangle_{\mathbb{R}^2} \mathrm{d}\mu = 0.$$

By the definition of the subgradient we have

$$f(z) + f^*(z^*) \geq \langle z | z^* \rangle_{\mathbb{R}^2} \text{ for all } z, z^* \in \mathbb{C}.$$

This shows that the above integrand is always non-negative. Hence, the integral is zero if and only if the integrand is zero almost everywhere. This implies $f(u(x)) + f^*(v(x)) = \langle u(x) | v(x) \rangle_{\mathbb{R}^2}$ and hence $v(x) \in \partial f(u(x))$ for almost every x. □

Now the subgradient of the L^1-norm is given by a simple corollary.

Corollary 3.3 (Subgradient of the L^1 norm). *The subgradient of Φ at v is given by*

$$\partial\Phi(v) = \operatorname{sign}(v).$$

The sign for L^2 functions is defined by

$$\operatorname{sign}(v) = \left\{ w \in L^2(\Omega, d\mu) \,\middle|\, w(x) \in \operatorname{sign}(v(x)) \text{ for almost every } x \right\}.$$

and the complex sign is given by

$$\operatorname{sign}(z) = \begin{cases} \left\{ \frac{z}{|z|} \right\} & \text{for } z \neq 0 \\ \{ \zeta \in \mathbb{C} | \, |\zeta| \leq 1 \} & \text{for } z = 0. \end{cases}$$

Furthermore $\operatorname{dom} \partial\Phi$ *is dense in* $L^2(\Omega, d\mu)$.

Proof. The formula for the subgradient is clear by the above lemma and the fact that $\partial |z| = \operatorname{sign} z$.

To show that $\operatorname{dom} \partial\Phi$ is dense, we mention that for every function $v \in L^2(\Omega, \mathrm{d}\mu)$ with finite support we have

$$\operatorname{sign} v \chi_{\operatorname{supp} v} \in \partial\Phi(v).$$

The functions with finite support are dense in $L^2(\Omega, \mathrm{d}\mu)$. □

3.1.2 Pointwise shrinkage

With the results of the previous subsection we are almost able to prove the main result of this section. We are going to show, that the subgradient descent along the $L^1(\Omega, \mathrm{d}\mu)$ norm in $L^2(\Omega, \mathrm{d}\mu)$ is pointwise soft shrinkage of the initial value.

To do so we need two theorems from the theory of semi groups which can be found in [Bre73]. The first theorem shows under what conditions a subgradient descent problem has a solution.

Theorem 3.4 (Existence for solutions of subgradient descents). *Let Ψ be a proper, convex and lower semi continuous functional on a real Hilbert space H. Then for every function $f \in \overline{\operatorname{dom} \partial\Psi}$ there exists a solution of the subgradient descent problem*

$$\partial_t u + \partial\Psi(u) \ni 0, \ u(0) = f$$

with $u(t) \in \operatorname{dom} \partial\Psi$ for all $t > 0$.

To solve descent problems numerically (as well as analytically) one can use discrete backward differences in time.

For a given $t > 0$ we fix $N \in \mathbb{N}$ and choose the time step $\Delta t = t/N$. We write $u^n_{\Delta t}$ for the approximation to the true solution u at time $n\Delta t$. Further we set $u^0_{\Delta t} = f$. Then the discretized descent equation is

$$\frac{u^n_{\Delta t} - u^{n-1}_{\Delta t}}{\Delta t} + \partial\Psi(u^n_{\Delta t}) \ni 0, \text{ for } n = 1, \dots, N$$

or equivalently

$$u^{n-1}_{\Delta t} \in (\mathrm{Id} + \Delta t \partial\Psi)(u^n_{\Delta t}), \text{ for } n = 1, \dots, N.$$

The second theorem we need states the invertibility of $(\mathrm{Id} + \Delta t \partial\Psi)$ and the convergence of the approximation towards the true solution.

Theorem 3.5 (Convergence of the difference approximation). *For $u \in H$ there exists a unique $v \in \operatorname{dom} \partial\Psi$ with*

$$u \in v + \Delta t \partial\Psi(v).$$

This function v is the minimizer of $\frac{1}{2\Delta t} \|v - u\|^2_H + \Psi(v)$.

If $u : [0, T] \to L^2(\Omega, d\mu)$ is a solution of the subgradient descent equation $\partial_t u + \partial\Phi(u) \ni 0$ for $T > 0$ it holds for every $0 < t < T$

$$\lim_{\substack{n \to \infty \\ n\Delta t = t}} u^n_{\Delta t} = u(t)$$

with $u^n_{\Delta t}$ defined as above.

The proof of this theorem can also be found in [Bre73]. Now we have everything we need to prove the main theorem of this section.

Theorem 3.6 (Descent along the L^1 norm). *Let $v_0 \in L^2(\Omega, d\mu)$. Then the solution of the descent equation*

$$\partial_t v + \partial \|v\|_{L^1(\Omega,d\mu)} \ni 0, \quad v(0) = v_0$$

is given by

$$(v(t))(x) = S_t(v_0(x)).$$

Proof. For a function $v \in L^2(\Omega, \mathrm{d}\mu)$ it holds

$$v(x) - S_{\Delta t}(v(x)) = \begin{cases} \operatorname{sign}(v(x))\Delta t, & |v(x)| > \Delta t \\ v(x), & |v(x)| \leq \Delta t. \end{cases}$$

We define the abbreviation $(T_{\Delta t}v)(x) = S_{\Delta t}(v(x))$ and together with Corollary 3.3 this yields

$$v - T_{\Delta t}(v) \in \Delta t \partial \|T_{\Delta t}(v)\|_{L^1(\Omega,\mathrm{d}\mu)}.$$

Theorem 3.5 states that

$$v(t) = \lim_{\substack{n\to\infty \\ n\Delta t = t}} T_{\Delta t}^n v_0$$

converges to a solution of our descent problem.

We observe that for $n\Delta t = t$ we have $T_{\Delta t}^n = T_t$, i. e. the sequence $T_{\Delta t}^n v_0$ is a constant sequence and therefore it converges to $(T_t v_0)(x) = S_t(v_0(x)) = (v(t))(x)$. □

3.1.3 Shrinkage after an isometrical transformation

The previous subsection showed, how pointwise soft shrinkage can be seen as a subgradient descent. But typically, shrinkage is applied after some transformation. The following lemma makes the results of the previous subsection applicable for this case.

Lemma 3.7 (Pullback of subgradients). *Let H_1 and H_2 be two Hilbert spaces, $A : H_1 \to H_2$ an invertible isometrical transformation, $F : H_2 \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous functional with dense* $\operatorname{dom} F$ *and define $G(u) = F(Au)$.*

For $u_0 \in H_1$ let u denote the solution of

$$\partial_t u + \partial G(u) \ni 0, \quad u(0) = u_0 \text{ in } H_1$$

and v denote the solution of

$$\partial_t v + \partial F(v) \ni 0, \quad v(0) = Au_0 \text{ in } H_2.$$

Then $v(t) = Au(t)$.

Proof. The functional G is proper, convex and lower semi continuous on H_1 since A is a linear isometry.

Now we calculate the subgradient of G:

$$\begin{aligned} w \in \partial G(u) &\Leftrightarrow G(u) + \langle w |\, v-u\rangle_{H_1} \leq G(v)\ \forall v \in H_1 \\ &\Leftrightarrow F(Au) + \langle Aw |\, Av - Au\rangle_{H_2} \leq F(Au)\ \forall Av \in \operatorname{range} A = H_2 \\ &\Leftrightarrow Aw \in \partial F(Au) \\ &\Leftrightarrow w \in A^{-1}\partial F(Au). \end{aligned}$$

This shows $\partial G(u) = A^{-1}\partial F(Au)$. Finally

$$\begin{aligned} & \partial_t u + \partial G(u) \ni 0, && u(0) = u_0 \\ \Leftrightarrow\ & \partial_t u + A^{-1}\partial F(Au) \ni 0, && u(0) = u_0 \\ \Leftrightarrow\ & \partial_t (Au) + \partial F(Au) \ni 0, && u(0) = u_0 \\ \Leftrightarrow\ & \partial_t v + \partial F(v) \ni 0, && v(0) = Au_0 \end{aligned}$$

where $v(t) = Au(t)$. □

In other words one can say, that the solutions of the two descent equations in H_1 and H_2 are connected by the operator A. The solution of the descent along G in H_1 is mapped by A on a solution of the descent along F in H_2.

3.2 Applications to different shrinkage methods

This section shows the application of the established results to different shrinkage methods. We are going to show, how to apply the results to various shrinkage methods, namely shrinkage of the discrete wavelet coefficients Fourier coefficients, the discrete Fourier coefficients and the continuous Fourier transform.

With the help of the results of the previous section, it is quite clear how to do this. Lemma 3.7 and Theorem 3.6 are everything we need. The isometrical transformation A in Lemma 3.7 is replaced by the transformations mentioned above and we have to specify the two Hilbert spaces H_1 and H_2 and apply the Theorem 3.6. This will be described in detail below.

Two cases cause a little trouble and can not be handled in this framework: The continuous wavelet transform and the stationary wavelet transform. This is because the transformations are not invertible isometries as needed for Lemma 3.7.

Figure 3.2: The image `eye`.

The case of the continuous wavelet transform was first described in [BLM03] and we present the results in Subsection 3.2.4. It turns out, that soft shrinkage of the continuous wavelet transform is not the solution of a descent equation.

In the case of the stationary wavelet transform Chambolle and Lucier showed, that the soft shrinkage can be interpreted as the solution of a different descent equation [CL01]. Further, Mrazek et al. obtained an equivalence result which is totally different from the results obtained here (see [MWS03]). It relates translation invariant wavelet shrinkage with the Haar wavelet to certain diffusion equations. We present this result briefly in Section 3.3.

Further we illustrate the effect of wavelet and Fourier shrinkage in the Subsection 3.2.5. We use the image `eye` which is a closeup on a man's eye. It is a suitable image for illustrative purposes because it provides very different regions: small and sharp details like the eyelashes, texture-like parts of different contrast like the eyebrows or the skin below the eye, smooth parts like the eyeball or the skin above the eye and sharp edges like the edge of the lower lid. Because this image provides many different features, we will use it thoughout the thesis. This has the additional effect that the results of different methods can be compared better.

The image has a resolution of 256 times 256 pixels, 256 gray levels and is shown at a resolution of 150 dots per inch in Figure 3.2. For calculations the gray levels have been scaled to the interval $[0, 1]$. It is highly recommended, to have a look at the illustrations digitally, for example on the authors website, because of better contrast and better scaling possibilities.

3.2.1 Discrete wavelet shrinkage

First we treat the case of discrete wavelet shrinkage. This is the place where shrinkage methods have their origin and where they are used the most.

We consider an orthogonal periodic wavelet base $\{\psi_\gamma\}_{\gamma\in\Gamma}$ of $L^2(I)$ as constructed in Subsection 1.2.1.

We define the orthogonal mapping $W : L^2(I) \to \ell^2(1 \cup \Gamma)$ via

$$f \mapsto \left(\langle f|\, 1\rangle, (f_\gamma)_{\gamma\in\Gamma}\right).$$

The mapping W is invertible and isometrical. We define Φ on $\ell^2(1 \cup \Gamma)$ by $\Phi(a) = \|a\|_{\ell^1(1\cup\Gamma)}$ (resp. ∞ if the series does not converge) and the functional we need is $\Psi : L^2(I) \to \overline{\mathbb{R}}$ defined by

$$\Psi(f) = \Phi(Wf) = \begin{cases} |\langle f|\, 1\rangle| + \|(f_\gamma)_{\gamma\in\Gamma}\|_{\ell^1(\Gamma)} & \text{whenever it exists} \\ \infty & \text{else.} \end{cases}$$

We obtain the result, that the discrete wavelet shrinkage

$$u(t) = S_t(\langle f|\, 1\rangle) + \sum_{\gamma\in\Gamma} S_t(f_\gamma)\psi_\gamma$$

is a solution of the descent equation

$$\partial_t u + \partial\Psi(u) \ni 0, \quad u(0) = f.$$

This result has a very nice interpretation. As we have seen in Subsection 2.2.4, the functional

$$\Psi(f) = |\langle f|\, 1\rangle| + \sum_{\gamma\in\Gamma} |f_\gamma|$$

is exactly the Besov norm of f in the Besov space $B^{d/2}_{1,1}(I)$. So the wavelet shrinkage of a d-dimensional signal is a descent along this Besov norm.

Remark 3.8. *Our result is slightly different from classical results about wavelet shrinkage. The usual wavelet shrinkage does not change the average $\langle f|\, 1\rangle$. By similar computations one can show the classical result [CL01], that the shrinkage*

$$u(t) = \langle f|\, 1\rangle + \sum_{\gamma\in\Gamma} S_t(f_\gamma)\psi_\gamma$$

is the solution of the subgradient descent along the Besov semi norm

$$\partial_t u + \partial\,|u|_{B^{d/2}_{1,1}(I)} \ni 0, \quad u(0) = f.$$

This wavelet shrinkage is illustrated in Figure 3.3.

3.2.2 Discrete Fourier shrinkage

At first we give a short explanation of the discrete Fourier transform (or the Fourier series expansion) of functions $f \in L^2(I)$. It is a well know fact that the family $e_k(x) = e^{2\pi i \langle k|x\rangle_{\mathbb{R}^d}}$, $k \in \mathbb{N}^d$ is an orthonormal basis of $L^2(I)$ (see for example [Rud87]).

Every element $f \in L^2(I)$ has a convergent expansion

$$f = \sum_{k\in\mathbb{N}^d} \langle f | e_k\rangle_{L^2(I)} e_k$$

which is called the Fourier series of f.

The discrete Fourier shrinkage of f is given by

$$u(t) = \sum_{k\in\mathbb{N}^d} S_t(\langle f | e_k\rangle_{L^2(I)}) e_k.$$

We define the mapping $\mathcal{F} : L^2(I) \to \ell^2\left(\mathbb{N}^d\right)$ via

$$f \mapsto (\langle f | e_k\rangle_{L^2(I)})_{k\in\mathbb{N}^d}$$

and further the functional $\Phi : \ell^2\left(\mathbb{N}^d\right) \to \overline{\mathbb{R}}$ via

$$\Phi(a) = \begin{cases} \sum_{k\in\mathbb{N}^d} |a_k| & \text{whenever it exists} \\ \infty & \text{else.} \end{cases}$$

Then the function which will do the desired job is

$$\Psi(f) = \Phi(\mathcal{F}(f)) = \sum_{k\in\mathbb{N}^d} \left|\langle f | e_k\rangle_{L^2(I)}\right|.$$

Theorem 3.6 together with Lemma 3.7 show that the discrete Fourier shrinkage of $f \in L^2(I)$ is a solution of the descent equation

$$\partial_t u + \partial\Psi(u) \ni 0, \quad u(0) = f.$$

Unfortunately this functional does not have a nice interpretation as the functional for the discrete wavelet shrinkage in the sense that Ψ is a norm in a well known Banach space.

Further we can make the same remark as for the discrete wavelet shrinkage: It is not useful to shrink the mean value (which is $\langle f | e_0\rangle$ in this case)

and one should omit it in the shrinkage procedure. One can show, that the shrinkage

$$u(t) = \langle f | e_0 \rangle + \sum_{k \neq 0} S_t(\langle f | e_k \rangle) e_k$$

which leaves the mean value unchanged is the descent along the functional $\Psi(f) = \sum_{k \neq 0} |\langle f | e_k \rangle|$.

In Figure 3.4 the effect of Fourier shrinkage is illustrated.

3.2.3 Continuous Fourier shrinkage

The case of the continuous Fourier transform now is easy. The continuous Fourier transform of a signal $f \in L^2\left(\mathbb{R}^d\right)$ is

$$\mathcal{F}f(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-i\langle \omega | x \rangle_{\mathbb{R}^d}} \mathrm{d}x.$$

Remark 3.9. *The upper formula for the Fourier transform is not well defined for function in $L^2\left(\mathbb{R}^d\right)$. We omitted the correct introduction of the Fourier transform on the space $L^2\left(\mathbb{R}^d\right)$ and refer to [Rud87].*

The Fourier transform is an invertible isometry from $L^2\left(\mathbb{R}^d\right)$ onto itself. The inversion formula is

$$\mathcal{F}^{-1}f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathcal{F}f(\omega) e^{i\langle \omega | x \rangle_{\mathbb{R}^d}} \mathrm{d}\omega.$$

We define the continuous Fourier shrinkage as

$$u(t) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} S_t(\mathcal{F}f(\omega)) e^{i\langle \omega | x \rangle_{\mathbb{R}^d}} \mathrm{d}\omega$$

and then it is clear, that it is the solution of the subgradient descent along the functional

$$\Psi(f) = \begin{cases} \|\mathcal{F}f\|_{L^1(\mathbb{R}^d)} & \text{whenever it exists} \\ \infty & \text{else.} \end{cases}$$

The continuous Fourier shrinkage is not further illustrated, because the continuous Fourier transform is usually implemented by the fast Fourier transform which approximates both the Fourier series expansion and the continuous Fourier transform. Thus the pictures would look like in Figure 3.4.

3.2.4 Continuous wavelet shrinkage

In the case of the continuous wavelet transform we get into trouble. However, the wavelet transform is an isometry

$$L^2(\mathbb{R}) \to \operatorname{range} L_\psi \subset L^2\left(\mathbb{R}^2, \mathrm{d}a\mathrm{d}b/a^2\right)$$

but the range of the wavelet transform is a proper subspace of the space $L^2\left(\mathbb{R}^2, \mathrm{d}a\mathrm{d}b/a^2\right)$. In particular $\operatorname{range} L_\psi$ is not invariant under shrinkage. Unfortunately, a subgradient descent in such a subspace is in general not a subgradient descent in the original Hilbert space and vice versa.

In the cases were the transformations A were onto, we used, that for two functionals F and G related by $G(u) = F(Au)$ the subgradients satisfy the equation $\partial G(u) = A^{-1}\partial F(Au)$.

So the equation for the subgradient descent along G is

$$\partial_t u + A^{-1}\partial F(Au) \ni 0.$$

We want to answer the question, what this equation means for the case of continuous wavelet shrinkage, i. e. where A is replaced by the operator L_ψ and the inverse of A is replaced by L_ψ^* which inverts L_ψ only on its range. Further we define as usual $\Phi : L^2\left(\mathbb{R}^2, \mathrm{d}a\mathrm{d}b/a^2\right) \to \overline{\mathbb{R}}$ defined by

$$\Phi(v) := \int_{\mathbb{R}^2} |v(a,b)| \, \frac{\mathrm{d}a\mathrm{d}b}{a^2}.$$

To analyze the descent equation

$$\partial_t u + L_\psi^* \partial\Phi(L_\psi u) \ni 0, \quad u(0) = f$$

we first examine $L_\psi^*\partial\Phi(L_\psi u)$.

The domain of $\partial\Phi$ is dense in $L^2\left(\mathbb{R}^2, \mathrm{d}a\mathrm{d}b/a^2\right)$ but it intersects the range of the wavelet transformation only at one point. To show this, we consider the subdifferential of a ϕ at a wavelet transformed function. The subdifferential $\partial\Phi(L_\psi f)$ is only nonempty, if $L_\psi f$ has integrable support (compare Corollary 3.3). However the support of a wavelet transformed function has finite measure if and only if the function is zero almost everywhere, see [Wil00].

Now we want to find a $g \in L^2(\mathbb{R})$ such that for a given $h \in L^2(\mathbb{R})$ the inclusion $h - g \in \Delta t L_\psi^* \partial\Phi(L_\psi g)$ is true (compare Theorem 3.5 and Theorem 3.6). The only possibility is $g = 0$ since for every other g we have $L_\psi^*\partial\Phi(L_\psi g) = \emptyset$.

Hence for any initial value $u^0_{\Delta t} = f$ and the sequence $u^n_{\Delta t}$ defined by

$$u^n_{\Delta t} - u^{n+1}_{\Delta t} \in \Delta t L^*_\psi \partial\Phi(L_\psi u^{n+1}_{\Delta t})$$

we obtain $\lim_{\substack{n\to\infty \\ n\Delta t = t}} u^n_{\Delta t} = 0$ if it exists. This shows, that this construction, which is in analogy to the construction of the solution for the other descent methods, has nothing to do with wavelet shrinkage.

3.2.5 Illustration

Here we illustrate effects of wavelet and Fourier shrinkage. Figure 3.3 shows the soft shrinkage of the discrete wavelet coefficients. The used wavelet is the `coif4` wavelet (as a compromise between smoothness and filter length). The wavelet decomposition of the input image is performed up to the fourth level. The shrinkage is performed only on the detail coefficients and the approximation coefficients are left unchanged.

Figure 3.4 shows the effect of the Fourier shrinkage. The effect can be compared with the classical linear low pass filtering. To make the images comparable, the threshold λ is chosen such that the number of non-zero Fourier coefficients is the same for the low pass filter and the soft Fourier shrinkage. One sees, that low pass filtering leads to more ringing artifacts near the edges and worse denoising in homogeneous regions. The low pass filter has the well known effect, that the noise is not removed completely and looks like "low frequency noise" after denoising. One advantage of the low pass filtering is, that some details are preserved better (like the eyelashes, which are removed by Fourier shrinkage).

Compared to the wavelet shrinkage one observes, that the denoising with either low pass filtering or Fourier shrinkage leads to more global artifacts. The artifacts from wavelet shrinkage are much more local which is due to the locality of the wavelet basis function in contrast to the Fourier basis functions.

Furthermore one observes, that wavelet shrinkage preserves more details than Fourier shrinkage. This effect is most clear for eyelashes which are totally deleted by Fourier shrinkage. The denoising of homogeneous regions is much better for the wavelet shrinkage (again du to the locality of the wavelets). Also the edges are preserved better by wavelet shrinkage than by Fourier shrinkage.

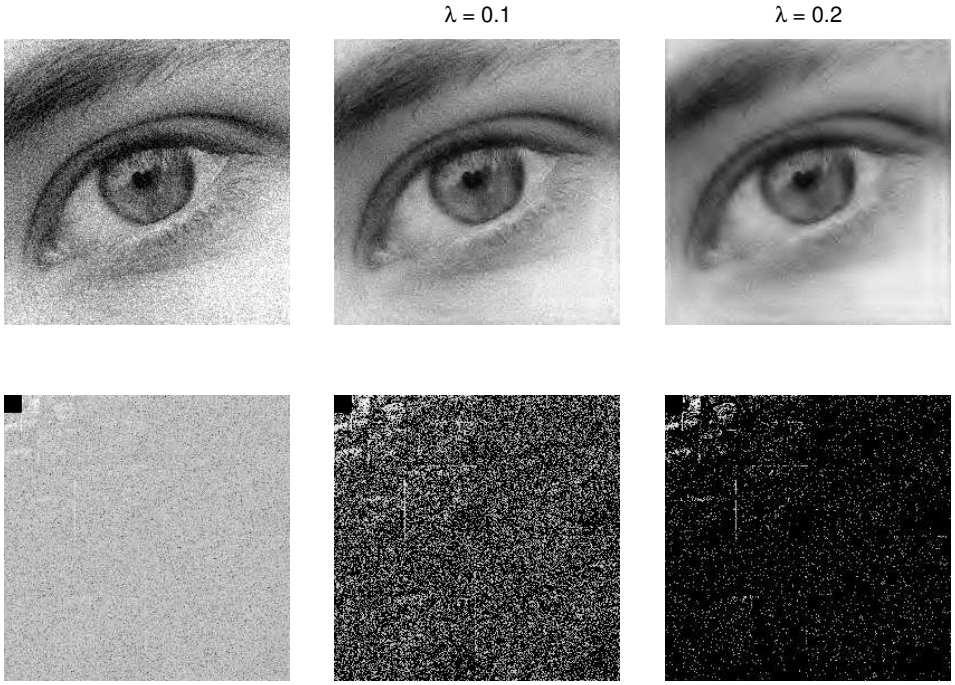

Figure 3.3: Illustration of discrete wavelet shrinkage. The used wavelet is the `coif4` wavelet. Top row, from left right: the noisy image, wavelet shrinkage for different values of λ. The second row shows the discrete wavelet transform of the upper row. The approximation coefficients are not shown because they are not changed by wavelet shrinkage.

3.3 Translation invariant wavelet shrinkage and diffusion equations

One drawback of discrete wavelet shrinkage is it dependence on translations. Coifman and Donoho presented in [CD95] a translation invariant wavelet shrinkage which is based on the wavelet transform of different translations of the signal. In this section we present the results of [MWS03] which say that (in a special case) the translation invariant wavelet shrinkage can be seen as the solution of a discrete version of a diffusion equation.

3.3.1 Translation invariant wavelet shrinkage and the stationary wavelet transform

In this subsection we only deal with discrete one dimensional signals, i. e. we think of time series rather than continuous signals.

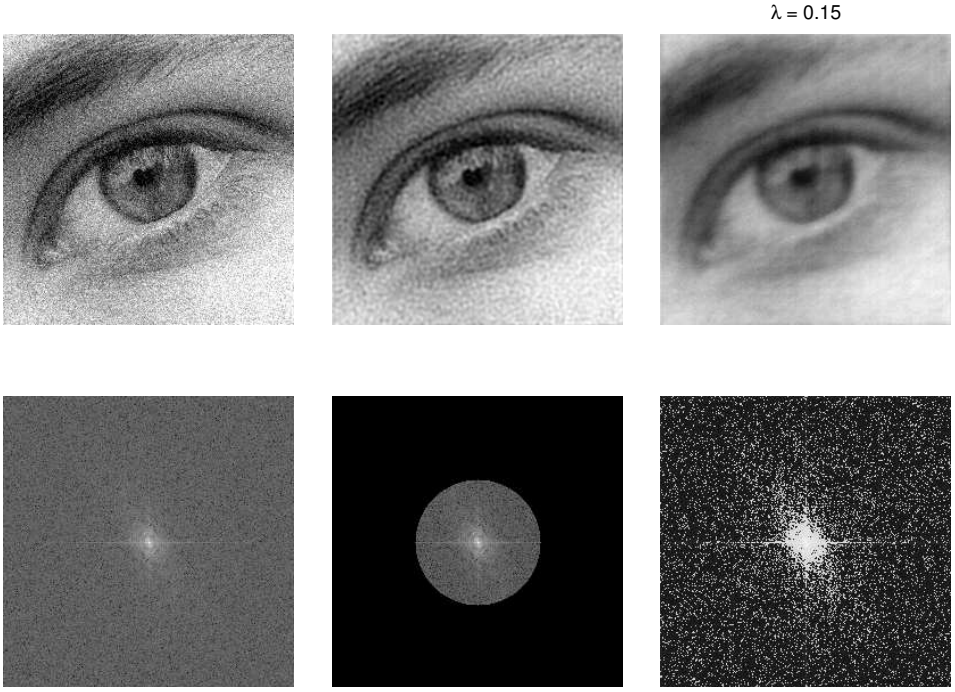

Figure 3.4: Illustration of Fourier shrinkage. Top row, from left right: the noisy image, low pass filtered image, Fourier shrinkage with $\lambda = .15$, The second row shows the Fourier transform of the upper row.

We consider a discrete signal $(f_i)_{i\in\mathbb{Z}}$. The idea behind translation invariant shrinkage is simple: Do not use only the signal itself but also translations of the signal. One transforms the translations (f_{i-n}) of the signal for different n, performs wavelet shrinkage for the translated signals, transforms the results back and finally averages over the different results.

Another way to calculate the translation invariant shrinkage is to use the so called stationary wavelet transform.

Definition 3.10 (Stationary wavelet transform). *Let $(f_i)_{i\in\mathbb{Z}}$ be a discrete signal and $h, \tilde{h}, g, \tilde{g}$ be the low resp. high pass filters of the wavelet analysis resp. reconstruction.*

The stationary wavelet transform of f at level J are the approximation coefficients c^J and detail coefficients d^j at level $j = 1, \ldots, J$ defined recursively by

$$c^0 = f$$

and

$$c^{j+1} = c^j * (h \uparrow 2^j) \qquad d^{j+1} = c^j * (g \uparrow 2^j) \qquad \text{for } 0 \le j < J.$$

The stationary wavelet transform obeys a simple inversion formula.

Proposition 3.11 (Reconstruction formula of the stationary wavelet transform). *The reconstruction formula for the stationary wavelet transform is*

$$c^j = \frac{1}{2}\left(c^{j+1} * (\tilde{h} \uparrow 2^j) + d^{j+1} * (\tilde{g} \uparrow 2^j)\right).$$

The proof of this proposition can be found in [Mal99].

With the help of the stationary wavelet transform we can write down the translation invariant wavelet shrinkage of a single level decomposition in just one line:

$$u = \frac{1}{2}\left(f * h * \tilde{h} + S_\lambda(f * g) * \tilde{g}\right).$$

3.3.2 The Haar wavelet leads to diffusion

The equivalence result we are going to obtain only involves the translation invariant wavelet shrinkage of a one level decomposition with the Haar wavelet. The filters of the Haar wavelet are

$$h = \frac{1}{\sqrt{2}}(\dots, 0, 1, 1, 0, \dots) \qquad g = \frac{1}{\sqrt{2}}(\dots, 0, -1, 1, 0, \dots)$$
$$\tilde{h} = \frac{1}{\sqrt{2}}(\dots, 0, 1, 1, 0, \dots) \qquad \tilde{g} = \frac{1}{\sqrt{2}}(\dots, 0, 1, -1, 0, \dots).$$

The translation invariant wavelet shrinkage of a signal f of a single level decomposition with the Haar wavelet reads as follows:

$$u_k = \frac{1}{4}(f_{k-1} + 2f_k + f_{k+1}) + \frac{1}{2\sqrt{2}}\left(-S_\lambda\left(\frac{f_{k+1} - f_k}{\sqrt{2}}\right) + S_\lambda\left(\frac{f_k - f_{k-1}}{\sqrt{2}}\right)\right).$$

Because the filters of the Haar wavelet are simple difference filters (or in other words: finite difference approximation of derivatives) this shrinkage rule looks a little like a discretized version of a differential equation. In fact,

this is true. We sort the above equation in a different way to make this more clear:

$$\begin{aligned} u_k &= f_k + \frac{f_{k+1}-f_k}{4} - \frac{f_k - f_{k-1}}{4} \\ &\quad + \frac{1}{2\sqrt{2}}\left(-S_\lambda\left(\frac{f_{k+1}-f_k}{\sqrt{2}}\right) + S_\lambda\left(\frac{f_k - f_{k-1}}{\sqrt{2}}\right)\right) \\ &= f_k + \left(\frac{(f_{k+1}-f_k)}{4} - \frac{1}{2\sqrt{2}} S_\lambda\left(\frac{f_{k+1}-f_k}{\sqrt{2}}\right)\right) \\ &\quad - \left(\frac{(f_k - f_{k-1})}{4} - \frac{1}{2\sqrt{2}} S_\lambda\left(\frac{f_k - f_{k-1}}{\sqrt{2}}\right)\right). \end{aligned}$$

We introduce a new function g and a new variable Δt by

$$\Delta t\, g(|x|) = \frac{1}{4} - \frac{1}{2\sqrt{2}x} S_\lambda\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{4} - \frac{1}{2\sqrt{2}\,|x|} S_\lambda\left(\frac{|x|}{\sqrt{2}}\right).$$

For the second identity we assumed, that the shrinkage function is an odd function, i. e. we have $S_\lambda(x) = S_\lambda(|x|)\operatorname{sign} x$. This is a natural assumption which just says, that positive and negative coefficients are treated in the same way. For the rest of this section we assume this for the shrinkage functions. If we express the shrinkage rule in terms of the function g we get

$$\frac{u_k - f_k}{\Delta t} = (f_{k+1} - f_k) g(|f_{k+1} - f_k|) - (f_k - f_{k-1}) g(|f_k - f_{k-1}|)$$

The fraction on the left hand side can be interpreted as a difference in time (as the choice Δt suggests). The philosophy behind this is that we think of u as a version of f at a later time. One could write $\frac{u_k - f_k}{\Delta t} \approx \partial_t f$. In this spirit, the right hand side is a difference of differences in the spatial dimension and one could write $(f' g(|f'|))'$ instead. So the shrinkage rule corresponds to a discretization of the following differential equation

$$\partial_t u = (u' g(|u'|))'$$

with initial condition $u(0) = f$. This equation is well known in image processing where it has the form

$$\partial_t u = \operatorname{div}(g(|\nabla u|)\nabla u).$$

It is called the Perona-Malik diffusion equation and has been introduced by Perona and Malik in [PM90]. The Perona-Malik equation is a nonlinear variant of the heat equation and the function g is called diffusivity. The

diffusivity controls the speed of diffusion depending on the magnitude of the gradient. Usually, g is chosen such that it is equal to one for small magnitudes of the gradient and goes down to zero for large gradients. Hence the diffusion stops at positions where the gradient is large. These areas are considered as edges of an image. One can show, that special choices of g lead to edge preserving or even edge enhancing equations [PM90, Kee02]. Since the Perona-Malik equation is nonlinear the existence of a solution is not obvious. In fact there are no solutions under some assumptions on g and the initial value [Kic97, KK98]. Under a slight regularization Lions et al. proved the existence of solutions in [CLMC92, ALM92].

3.3.3 Shrinkage functions and diffusivities

Here we show how some properties of shrinkage functions and diffusivities are related. Further we are going to present corresponding diffusivity functions to some known and widely used shrinkage functions. The other way round we present shrinkage functions which are induced by famous diffusivities.

First we give an easy proposition which relates some properties of shrinkage functions and diffusivities. We formulate this relations for the case $\Delta t = 1/4$ which is a common choice and widely used for the Perona-Malik equation. This choice makes the relations more clear and relates some nice properties of the shrinkage function to other nice properties of the diffusivity. For example it leads to a normalized diffusion for reasonable shrinkage functions.

Proposition 3.12 (Corresponding properties of shrinkage functions and diffusivities). *Let $\Delta t = 1/4$. Then the diffusivity and the shrinkage function are related through*

$$g(|x|) = 1 - \frac{\sqrt{2}}{|x|} S_\lambda \left(\frac{|x|}{\sqrt{2}} \right).$$

It holds:

1. *If S_λ really shrinks the values then the diffusion is always forward, i. e.*

$$|S_\lambda(x)| \leq |x| \Leftrightarrow g(x) \geq 0.$$

2. *If S_λ is differentiable at 0 it holds: If the diffusion is equal to one for small gradients, the shrinkage function is flat around zero, i. e.*

$$g(x) \overset{x \to 0}{\to} 1 \;\Leftrightarrow\; S_\lambda(0) = 0 \text{ and } S'_\lambda(0) = 0.$$

3. *If the diffusion stops for large gradients the shrinkage function has linear growth at infinity, i. e.*

$$g(x) \overset{x\to\infty}{\to} 0 \;\Leftrightarrow\; \frac{S_\lambda(x)}{x} \overset{x\to\infty}{\to} 1.$$

Proof. The proof is an easy consequence of the relation between g and S_λ. □

Some examples will make the correspondence more clear. We choose $\Delta t = 1/4$ for the examples for the reasons explained above.

Example 3.13 (Linear shrinkage (damping)). *A linear shrinkage rule is*

$$S_\lambda(x) = \frac{x}{1+\lambda}$$

(compare to Subsection 4.1.1). The corresponding diffusivity is constant:

$$g(|x|) = \frac{\lambda}{(1+\lambda)}$$

So the diffusion is just linear and the speed of diffusion is $\lambda/(1+\lambda)$.

Example 3.14 (Soft shrinkage). *The well known soft shrinkage function $S_\lambda(x) = (|x| - \lambda)_+ \operatorname{sign} x$ gives*

$$g(|x|) = \left(1 - \frac{(|x| - \sqrt{2}\lambda)_+}{|x|}\right)$$

which is a diffusivity in the spirit of Perona and Malik. It is decreasing and according to the above proposition it holds $g(0) = 1$ and $g(x) \to 0$ for $x \to \infty$.

We should remark, that this diffusivity is a stabilized total variation diffusivity, i. e. it corresponds to the total variation flow (see [MWS03] for details).

Example 3.15 (Hard shrinkage). *The hard shrinkage function (or hard thresholding)*

$$S_\lambda(x) = x(1 - \chi_{[-\lambda,\lambda]}(x))$$

leads to

$$g(|x|) = \begin{cases} 1 & \text{for } |x| \le \sqrt{2}\lambda \\ 0 & \text{else.} \end{cases}$$

This is a "piecewise linear diffusion" where diffusion is forbidden if the derivative has absolute value larger that $\sqrt{2}\lambda$.

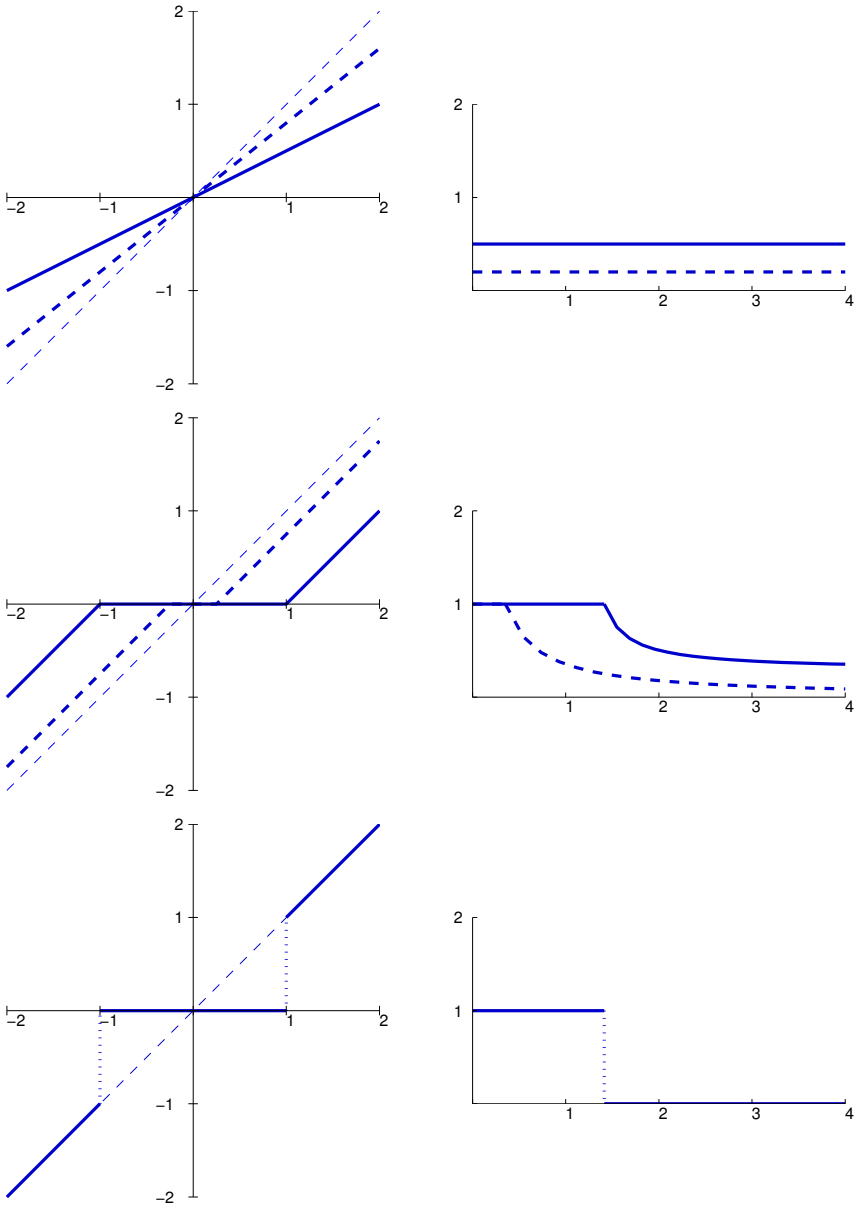

Figure 3.5: Shrinkage functions (left) and diffusivities (right). First row: Linear damping with $\lambda = 1$ (solid), and $\lambda = 1/4$ (dotted). Second row: Soft shrinkage with $\lambda = 1$ (solid), and $\lambda = 1/4$ (dotted). Third row: Hard shrinkage with $\lambda = 1$. The parameter Δt is set to $1/4$ for all plots. The thin dashed line is the diagonal.

These three examples are illustrated in Figure 3.5.

The other way round one can ask, how the shrinkage functions for famous diffusivities look like. The function S_λ expressed in terms of g looks like

$$S_\lambda(|x|) = |x|\,(1 - g(\sqrt{2}\,|x|)).$$

The dependence of the shrinkage function on the parameter λ will be fulfilled because usually diffusivities involve a parameter too.

Example 3.16 (Perona-Malik diffusivities). *Perona and Malik introduced in [PM90] two diffusivities*

$$g^1(x) = \frac{1}{1 + \frac{x^2}{\lambda^2}} \quad \textit{and} \quad g^2(x) = e^{-\frac{x^2}{\lambda^2}}.$$

The corresponding shrinkage functions for are

$$S^1_\lambda(x) = \frac{2x^3}{\lambda^2 + 2x^2} \quad \textit{and} \quad S^2_\lambda(x) = x\left(1 - e^{-\frac{2x^2}{\lambda^2}}\right)$$

and are illustrated in Figure 3.6.

More examples of shrinkage functions and diffusivities can be found in [MWS03]. In the work [MW03] Mrazek and Weickert extended the equivalence results to the two dimensional case. They proposed a translation and rotation invariant wavelet shrinkage scheme inspired by corresponding diffusion equations.

Illustrations for the effect of translation invariant wavelet shrinkage resp. nonlinear diffusion can be found in Figure 3.8. Because we developed the connection between diffusion equations and translation invariant wavelet shrinkage only in one dimension, the methods are applied to a one dimensional function. We use a variant of the test function `bumps` (see Figure 3.7) which has its origin in the paper [DJ94]. The function `bumps` which is used here is defined on the interval $I = [0, 1]$ by the formula

$$f(x) := \sum_{i=1}^{11} h_i k(\frac{x - a_i}{w_i}), \quad k(x) := \frac{1}{1 + t^2}$$

and the values for the heights h, the positions a and the widths w of the kernel k are

$$\begin{array}{rcl} h & = & [\,4 \quad 5 \quad 3 \quad 4 \quad 5 \quad 4.2 \quad 2.1 \quad 4.3 \quad 3.1 \quad 5.1 \quad] \\ a & = & [\,.1 \quad .13 \quad .15 \quad .23 \quad .25 \quad .4 \quad .44 \quad .65 \quad .76 \quad .78 \quad .81] \\ w & = & \frac{1}{40}[.05 \quad .05 \quad .06 \quad .1 \quad .1 \quad .3 \quad .1 \quad .1 \quad .05 \quad .08 \quad .05] \end{array}.$$

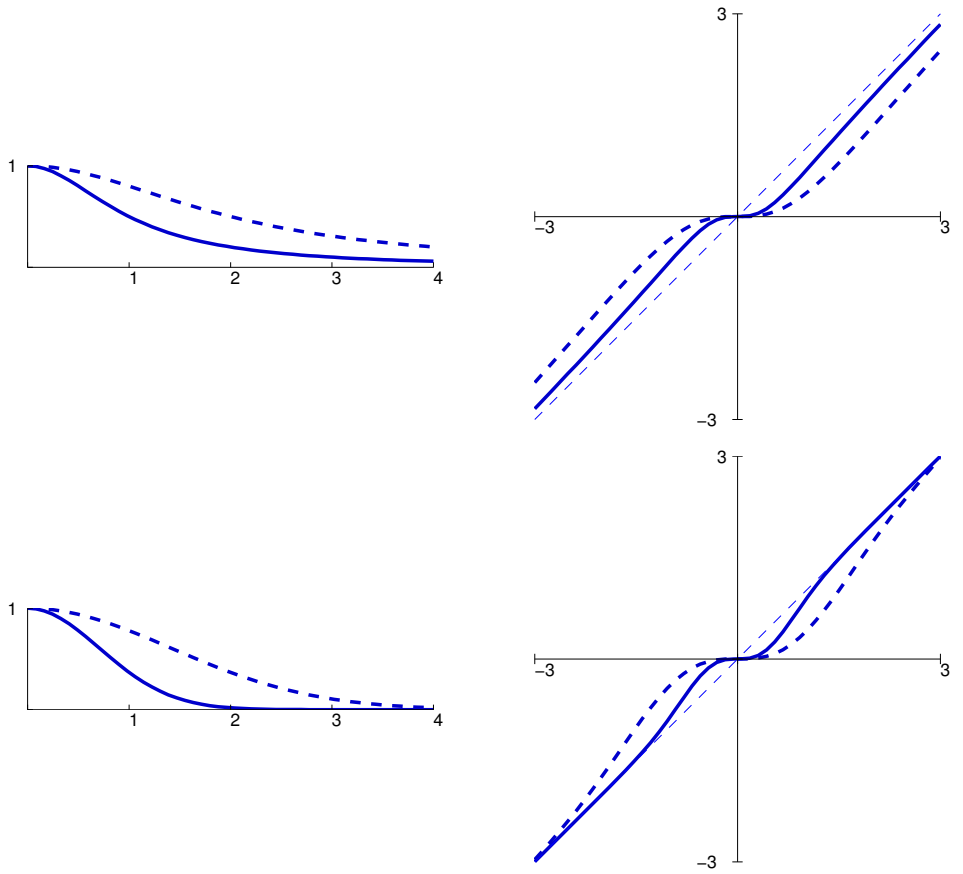

Figure 3.6: Diffusivities (left) and shrinkage functions (right). First row: Perona-Malik function g^1, $\lambda = 1$ (solid), and $\lambda = 2$ (dashed). Second row: Perona-Malik function g^2, $\lambda = 1$ (solid), and $\lambda = 2$ (dashed). The parameter Δt is set to $1/4$ for all plots. The thin dashed line is the diagonal and is plotted to make the plot easier to read.

The function f is sampled 1024 times on the interval I and disturbed by white noise of variance 0.2.

In Figure 3.8 one can see, that different choices of the shrinkage function have great influence on the denoising effect.

The shrinkage functions which are close to the identity for large values lead to less noise reduction but better preservation of the height of the bumps. Namely these are the hard shrinkage and the shrinkage which is related to the second Perona-Malik diffusivity.

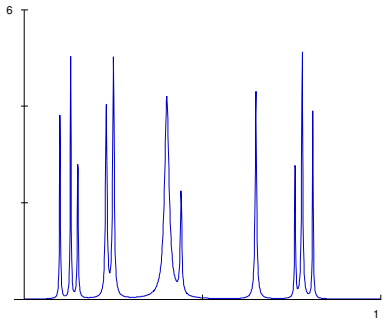

Figure 3.7: The test function `bumps`

The other shrinkage (damping and soft shrinkage) functions exhibit better denoising properties but the height of the bumps is changed much more. The first Perona-Malik diffusivity is a kind of compromise between these two cases.

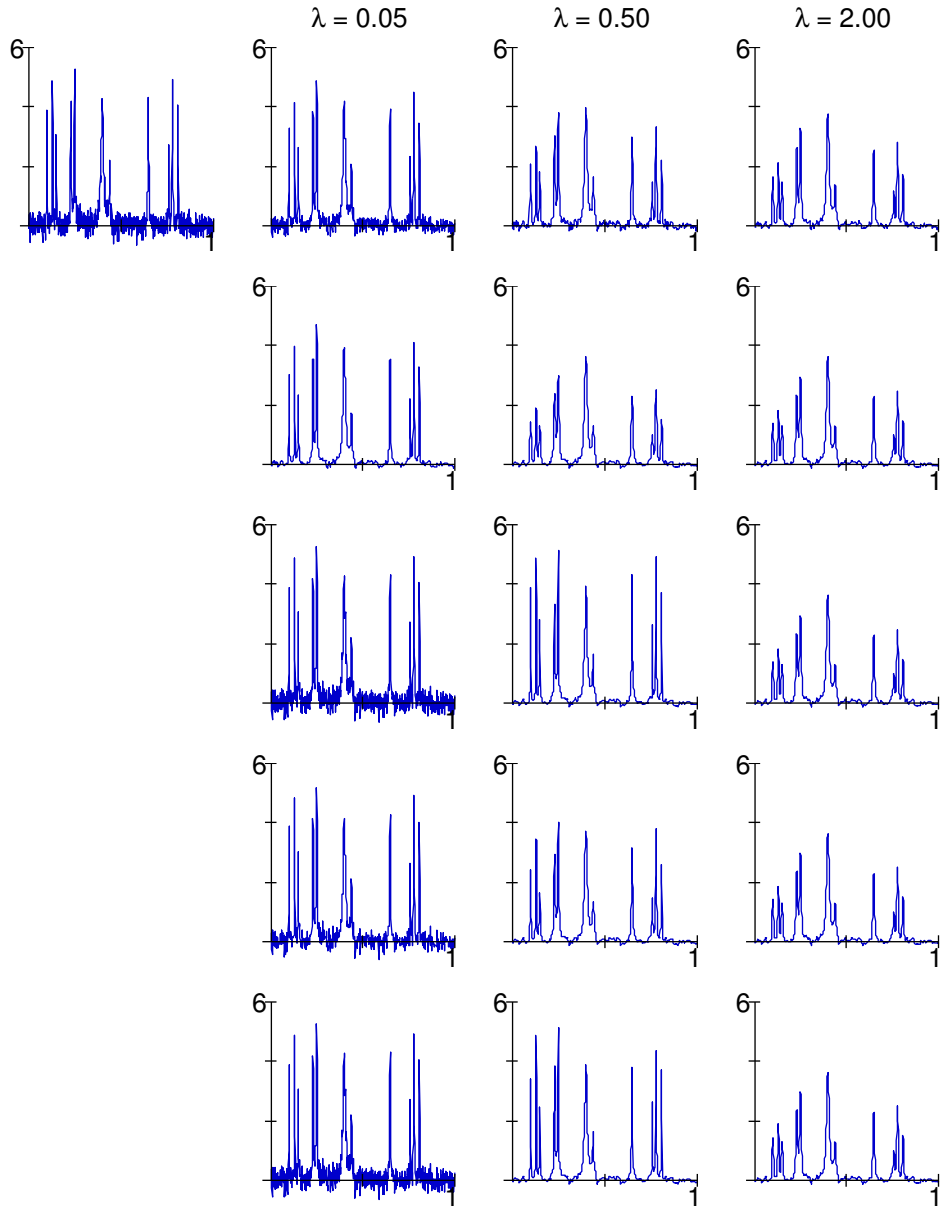

Figure 3.8: Denoising of a noisy signal with translation invariant wavelet shrinkage. Shrinkage is performed on the one level stationary wavelet decomposition with the Haar wavelet and iterated 15 times. Top left: Noisy signal. First row: linear damping as in Example 3.13, second row: soft shrinkage as in Example 3.14, third row: hard shrinkage as in Example 3.15, forth row: shrinkage function according to the first Perona-Malik diffusivity as in Example 3.16, fifth row: shrinkage function according to the second Perona-Malik diffusivity as in Example 3.16.

CHAPTER 4

Wavelet shrinkage and variational methods

In this chapter we collect results about variational methods which are related to wavelet shrinkage. By a variational method we mean, that the method is formulated as the minimization of a certain functional. For denoising the variational functional usually consists of two terms: a distance measure and a penalty term which penalizes "roughness". If f is a given noisy image a general functional for denoising looks like

$$\|u - f\|_X + \lambda \|u\|_Y$$

where $\|u - f\|_X$ measures the distance and $\|u\|_Y$ penalizes "roughness". The parameter λ is a weight parameter.

The first extensive treatment of variational methods with wavelets in image processing is [CDLL98].

Besov spaces are suited to measure smoothness, because they cover a wide range of smoothness classes. Further, we have seen in Section 2.2 that Besov norms can be expressed through wavelet coefficients. For this reasons we consider the following variational functionals:

$$\|u - f\|^2_{L^2(I)} + 2\lambda \, |u|^p_{B^s_{p,p}(I)}$$

or the slightly different variant

$$\|u - f\|^2_{L^2(I)} + 2\lambda \, |u|_{B^s_{p,p}(I)} .$$

The first case is treated in Section 4.1 and the second one is studied in Section 4.2.

Although the two functionals are almost the same, both are treated here in detail for two reasons: The first reason is, that they result in two different interpretations of shrinkage – the first one leads to different shrinkage functions which are applied to each wavelet coefficient and the second functional allows an interpretation of the minimizer as "global shrinkage" of f. The second reason is, that the two functionals allow the treatment of different limiting cases. In the first case it is possible to investigate the case $0 < p < 1$ and even $p \to 0$. However, the second case allows the analysis of $p = \infty$ which does not make sense in the first case. Both limiting cases are of interest because $p \to 0$ is related to hard shrinkage and $p = \infty$ corresponds to Hölder penalty terms. Other cases of interest are $p = 1, 2$, because for $p = 1$ the Besov spaces are related to the space BV of functions of bounded variation and for $p = 2$ they are equal to the Sobolev spaces $H^s(I)$.

The first variational functional and some special cases have been studied by DeVore, Lucier, Chambolle et al. in the works [DL92a, CDLL98, CL01]. We give a review of the results and some refinements which have been published in [Lor04b]. The results concerning the second variational functional are based on the work [Cha04] and can be found in [Lor04a].

The different methods will be illustrated with the signal `bumps` and the image `eye` and in the last section we present a short comparison of the different shrinkage methods obtained in chapter.

4.1 The scale of penalties $|\cdot|^p_{B^s_{p,p}(I)}$

Our first choice for the smoothness measure is the Besov semi norm $|\cdot|_{B^s_{p,p}(I)}$ raised to the power p. It is expressed through wavelet coefficients in the following way:

$$|f|^p_{B^s_{p,p}(I)} \asymp \sum_{\gamma\in\Gamma} 2^{sjp} 2^{jd/2(p-2)} |f_\gamma|^p .$$

This is probably the easiest choice one can make, because in this case the variational problem completely decouples. It has the following form: Find the coefficients u_γ that minimize

$$\sum_{\gamma\in\Gamma} \left(|u_\gamma - f_\gamma|^2 + 2\lambda 2^{sjp} 2^{jd/2(p-2)} |u_\gamma|^p \right).$$

Thus we can treat every wavelet coefficient for itself.

4.1.1 Shrinkage of single coefficients

Above we have seen that we can minimize our functional for every coefficient separately. If we want to know the coefficient u_γ of the minimizer u we just have to know the wavelet coefficient f_γ and not the whole set of coefficients. In this subsection we show how one can find this minimizing coefficient u_γ. This problem is equivalent to the construction of the so called *proximal mapping* in convex analysis. The proximal mapping of a mapping $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ is defined as

$$P_\lambda \phi(x) = \operatorname*{argmin}_w \left\{ (w-x)^2 + 2\lambda\phi(w) \right\}.$$

The proximal mapping is another example for a mapping which is in general multivalued.

The following lemma gives a description of the proximal mapping in terms of the subgradient.

Lemma 4.1 (Description of the proximal mapping in terms of the subgradient). *For any proper function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ and any $\lambda > 0$ one has*

$$P_\lambda \phi(x) \subset (\mathrm{Id} + \lambda \partial\phi)^{-1}(x).$$

If ϕ is convex, one has an equality.

Proof. Let $\phi_0(w) = (w-x)^2$. Then $P_\lambda \phi$ is given by the minimum point of $\phi_0 + 2\lambda\phi$. Because ϕ_0 is differentiable we have the necessary condition

$$-\phi_0'(w) \in 2\lambda \partial\phi(w)$$

for w belonging to $P_\lambda\phi(x)$ (which is sufficient if ϕ is convex). Since $\phi_0'(w) = 2(w-x)$ this condition can be written as

$$x \in (\mathrm{Id} + \lambda\partial\phi)(w)$$

which gives the assertion. □

To solve our problem we have to consider the weight functions $\phi(x) = |x|^p$. We are going to calculate the associated proximal mappings for this family of functions and we treat the following cases separately:

- $1 < p < \infty$
- $p = 1$
- $0 < p < 1$

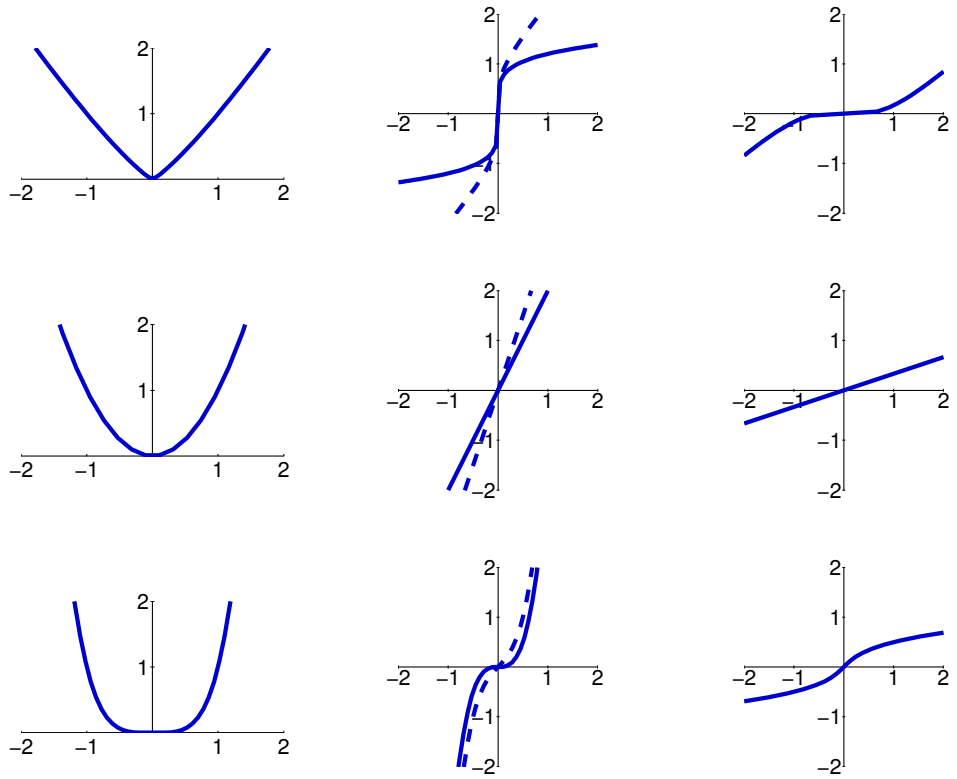

Figure 4.1: An illustration for the proximal mappings for $1 < p < \infty$. Upper row: $p = 1.2$, $\lambda = 1$. Mid row: $p = 2$, $\lambda = 1$. Lower row: $p = 4$, $\lambda = 1$. Left: $\phi(x) = |x|^p$, middle: solid $\partial\phi$, dashed $(\mathrm{Id} + \lambda\partial\phi)$, right: $P_\lambda\phi = (\mathrm{Id} + \lambda\partial\phi)^{-1}$.

The case $1 < p < \infty$

For $1 < p < \infty$ the function $\phi(x) = |x|^p$ is strictly convex and differentiable. The above lemma says, that the associated proximal mapping is the inverse of the function

$$x \mapsto x + \lambda p \operatorname{sign}(x) |x|^{p-1} .$$

For numerical purposes this function can be evaluated with a high precision with a few steps of Newton's method.

We should mention, that for the special case $p = 2$ we have the explicit formula $P_\lambda\phi(x) = \frac{1}{1+2\lambda}x$ which is a linear damping. An illustration of these proximal mappings is given in Figure 4.1.

The case $p = 1$

The case $p = 1$ is almost as easy as the case $p > 1$. The function $\phi(x) = |x|$ is not differentiable but still convex. The subgradient of ϕ is

$$\partial\phi(x) = \operatorname{sign} x := \begin{cases} -1 & , x < 0 \\ [-1, 1] & , x = 0 \\ 1 & , x > 0 \end{cases}$$

The function $(\mathrm{Id} + \lambda\partial\phi)$ is

$$(\mathrm{Id} + \lambda\partial\phi)(x) = \begin{cases} -x - \lambda & , x < 0 \\ [-\lambda, \lambda] & , x = 0 \\ x + \lambda & , x > 0 \end{cases}$$

and its inverse is

$$(\mathrm{Id} + \lambda\partial\phi)^{-1}(x) = \begin{cases} x + \lambda & , x \leq -\lambda \\ 0 & , x = 0 \\ x - \lambda & , x \geq \lambda \end{cases} .$$

Figure 4.2 illustrates this construction.

We see that the proximal mapping is in fact the soft shrinkage function. This is the second place, where the soft shrinkage function appears is a context totally different from denoising in image or signal processing: As the proximal mapping of the absolute value. This observation is the key to the famous equivalence result on wavelet shrinkage and variational denoising which was first presented in [CDLL98]. One could say that soft shrinkage is the natural non linear generalization of damping when one deals with "L^1" penalties instead of "L^2" penalties.

The case $0 < p < 1$

What happens in the case $0 < p < 1$? The function $\phi(x) = |x|^p$ is not convex anymore and the Lemma 4.1 only gives an inclusion of graphs. Furthermore ϕ is no more differentiable and even the subgradient at 0 degenerates: We have $\partial\phi(0) = \mathbb{R}$.

Figure 4.3 shows the same functions as Figure 4.1 and 4.2 for $p = 0.5$. We see that in this case the inverse function of $(\mathrm{Id} + \lambda\partial\phi)$ is multivalued. All we know is that the graph of the proximal mapping is a subset of the graph of $(\mathrm{Id} + \lambda\partial\phi)^{-1}$.

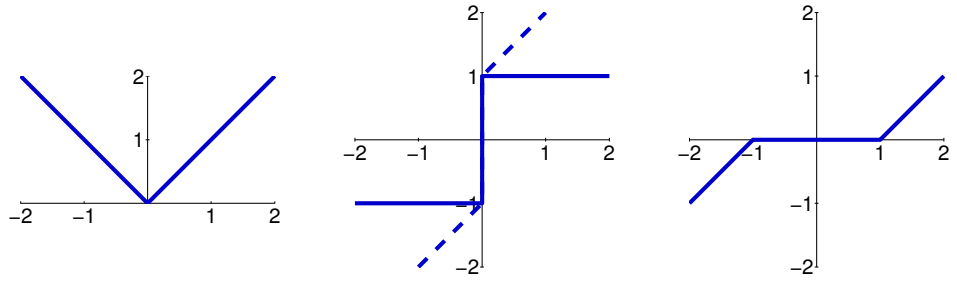

Figure 4.2: An illustration for the proximal mappings for $p = 1$, $\lambda = 1$. Left: $\phi(x) = 1$, middle: solid $\partial\phi(x) = \operatorname{sign} x$, dashed $(\mathrm{Id} + \lambda\partial\phi)$, right: $(\mathrm{Id} + \lambda\partial\phi)^{-1}$

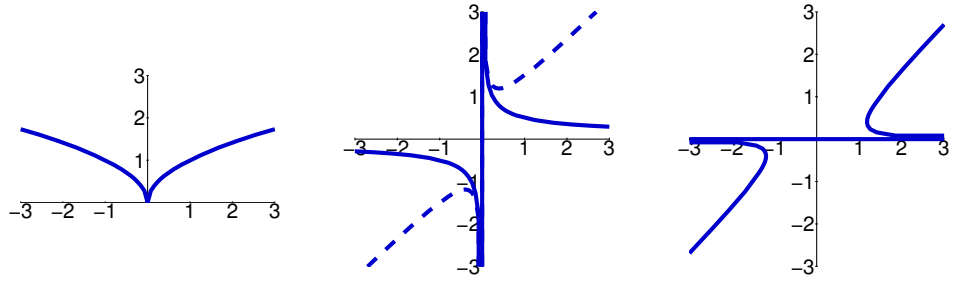

Figure 4.3: An illustration for the proximal mappings for $p = 0.5$, $\lambda = 1$. Left: $\phi(x) = \sqrt{|x|}$, middle: solid $\partial\phi(x) = \frac{\operatorname{sign} x}{2\sqrt{|x|}}$, dashed $(\mathrm{Id} + \lambda\partial\phi)$, right: $(\mathrm{Id} + \lambda\partial\phi)^{-1}$

To find the true proximal mapping $P_\lambda\phi$ have to take a closer look on the function

$$E_\lambda^x(w) = (w - x)^2 + 2\lambda\,|w|^p$$

and its minimizers. For simplification we assume $x, w > 0$ and we can forget about the absolute value. The other cases can be argued by symmetry. Before we make the exact calculations, we explore the dependence of the minimizers of E on the parameters x and λ "graphically".

In Figure 4.4 one can see, that for small x the function E_λ^x is monotone and the only minimum is the global minimum $w = 0$. If x increases, the minimum splits into two local minima and one local maximum. The value of the new local minimum is not necessarily smaller that $E_\lambda^x(0) = x^2$ and we have to figure out, which one is the global minimum. This "bifurcation" can also be observed in Figure 4.3 where we see three branches of the function $(\mathrm{Id} + \partial\phi)^{-1}$ for large values of x. The lower branch (which is constant 0) corresponds to the minimum of E_λ^x at 0 and the two upper branches are the

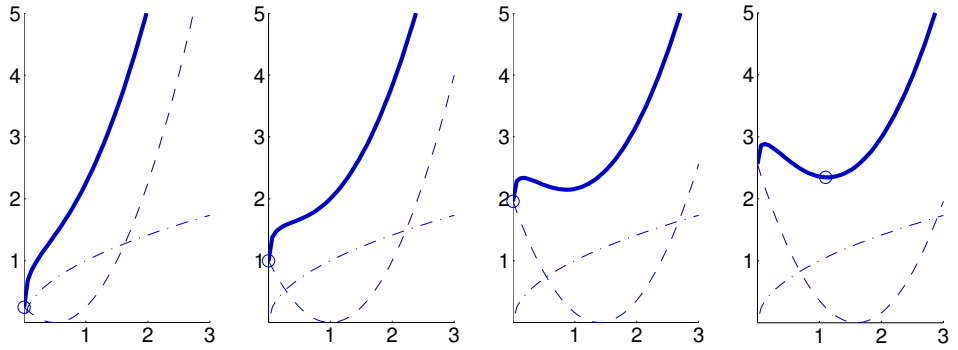

Figure 4.4: The function E^x_λ for $\lambda = 1$, $p = 1/2$ and different values of x. From left to right: $x = 0.5, 1, 1.4, 1.6$. Dashed: $w \mapsto (w-x)^2$, dash-dotted: $w \mapsto 2\lambda\sqrt{w}$, solid: $w \mapsto E^x_\lambda(w)$. The dots mark the global minimum of the function E^x_λ.

new local minimum and maximum. So our task is to find the point, where the proximal mapping jumps from the lower branch to the upper branch.

If this jumping point is known, the mapping $P_\lambda\phi$ can be evaluated easy: If $|x|$ is smaller than the jumping point, $P_\lambda\phi(x)$ is zero, else it can be evaluated with a few steps of Newton's method.

The idea to find this point is, to locate the points (x, w) where the two local minima have the same value, i. e. which satisfy $E^x_\lambda(w) = x^2$. In this equation we can write x as a function of w and get

$$x(w) = \frac{(2\lambda w^{p-2} + 1)w}{2}.$$

The point where the proximal mapping jumps from the zero branch to the upper branch is the minimum value of $x(w)$. This can be easily computed: The minimizer is

$$w_M = \Big(2\lambda(1-p)\Big)^{\frac{1}{2-p}},$$

and hence the jumping point is

$$x_J = x(w_M) = \frac{2-p}{2-2p}\Big(2\lambda(1-p)\Big)^{\frac{1}{2-p}}.$$

For $w < 0$ the situation is similar and we have the jumping point $-x_J$. We call this jumping point the "effective threshold" and denote it by λ_{eff}. We have to note, that the shrinkage function which is in fact the proximal mapping, is set valued. At the jumping point λ_{eff} it takes the value 0 and

w_M. This causes no trouble in the implementations, if we decide always to choose the value 0.

We collect the results of this subsection in the following theorem.

Theorem 4.2 (Shrinkage functions for different penalties). *Let $\phi^p(x) = |x|^p$. The mapping $P_\lambda^p(x) := \operatorname{argmin}_{w\in\mathbb{R}} \left\{(w-x)^2 + 2\lambda\phi^p(w)\right\}$ is given by the following formulas:*

- *For $1 < p < \infty$ the mapping P_λ^p is the single valued inverse of the mapping $x \mapsto x + \lambda p \operatorname{sign}(x)|x|^{p-1}$.*
- *The mapping P_λ^1 is the soft shrinkage function*

$$P_\lambda^1(x) = (|x| - \lambda)_+ \operatorname{sign}(x).$$

- *For $0 < p < 1$ the mapping P_λ^p is*

$$P_\lambda^p(x) = \begin{cases} 0 & \text{for } |x| \le \lambda_{\text{eff}} \\ \text{the value of largest absolute} & \\ \text{value of the inverse mapping} & \text{for } |x| \ge \lambda_{\text{eff}} \\ \text{of } x \mapsto x + \lambda p|x|^{p-1}\operatorname{sign}(x) & \end{cases}$$

where the effective threshold λ_{eff} is defined by $\lambda_{\text{eff}} = \frac{2-p}{2-2p}\Big(2\lambda(1-p)\Big)^{\frac{1}{2-p}}$.

We have to remark, that similar results have been obtained in [ML99] by Moulin and Liu. The author has not been aware of this work and found the results independently. Furthermore, in [ML99] the denoising problem is treated in a Bayesian framework (compare Chapter 5 where this point of view is briefly sketched). Here we developed the different shrinkage functions in a deterministic framework under the aspect of minimization of certain functionals which involve Besov penalty terms. Both ways lead to the same results, but from different starting points, and this gives further understanding of shrinkage methods and of the relations between different interpretations of shrinkage.

An illustration of the shrinkage functions and of the behavior of the effective threshold λ_{eff} can be found in Figure 4.5.

4.1.2 The limit $p \to 0$

In the previous subsection it is figured out, how the shrinkage functions corresponding to the weight functions $\phi(x) = |x|^p$ for $0 < p < \infty$. The limit $p \to 0$ is of interest, because it will lead us to the hard shrinkage function.

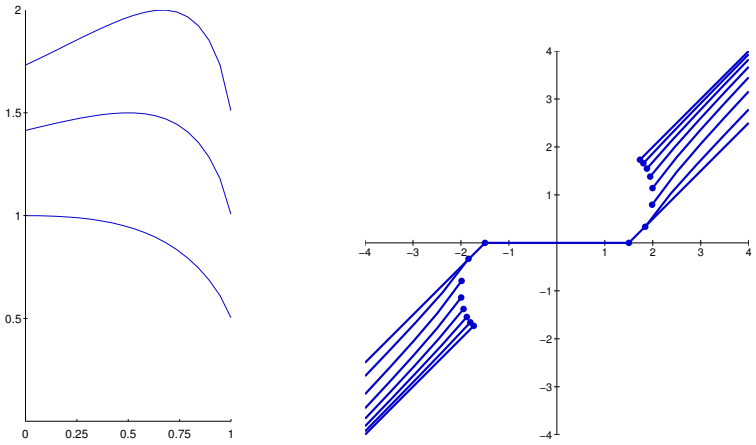

Figure 4.5: Illustration of the effective threshold and the interpolation between soft and hard thresholding. Left: The effective threshold λ_{eff} in dependence of $p \in [0,1]$ for $\lambda = 0.5, 1, 1.5$. Right: The shrinkage functions resp. proximal mappings P^p_λ from Theorem 4.2 associated to $\phi(x) = |x|^p$ for $p = 0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 1$ and $\lambda = 1.5$.

For $p \to 0$ the weight function degenerates to

$$\phi^0(x) = \begin{cases} 0 & , x = 0 \\ 1 & , x \neq 0. \end{cases}$$

The proximal mapping for this function is easy to calculate and it turns out, that we have

$$P^0_\lambda(x) := \begin{cases} x & , \ |x| \geq \sqrt{2\lambda} \\ 0 & , \ |x| \leq \sqrt{2\lambda}. \end{cases}$$

This function is known as the hard shrinkage or hard threshold function in signal and image denoising.

The hard shrinkage function is also the pointwise limit of the associated proximal mappings for $\phi(x) = |x|^p$ and $p \to 0$. For $0 < p < 1$ the proximal mappings interpolate between the soft and the hard shrinkage function (see Figure 4.5 for illustration).

4.1.3 Denoising with different shrinkage functions and illustrations

The results about shrinkage functions will be applied to the functionals

$$\|u - f\|^2_{L^2(I)} + 2\lambda \, |u|^p_{B^s_{p,p}(I)}$$

in this subsection. This corresponding denoising methods are applied to a variant of the standard test signal `bumps` as in Section 3.3. For all illustrations the wavelet `sym4` is used. By the results of the previous subsection it is clear that the minimizer of the above functional is obtained by applying the shrinkage functions from Theorem 4.2 with a level dependend threshold to the wavelet coefficients of f.

The cases $p = 2$, $p = 1$, $p = 1/2$ will be illustrated in detail here. The case $p = 1/2$ shall be seen only as an example for $0 < p < 1$. Further we are going to treat the limiting case $p \to 0$.

The penalty $|\cdot|^2_{B^s_{2,2}(I)}$, damping

The easiest penalty term probably is $|\cdot|^2_{B^s_{2,2}(I)}$ because it leads to a linear method. It is expressed in wavelet coefficients like

$$|f|^2_{B^s_{2,2}(I)} \asymp \sum_{\gamma\in\Gamma} 2^{2sj} |f_\gamma|^2 .$$

The minimization of the functional

$$\|u - f\|^2_{L^2(I)} + 2\lambda |u|^2_{B^s_{2,2}(I)}$$

leads to a simple scale dependent damping where the coefficients of the minimizer are given by

$$u_\gamma = \frac{1}{1 + 2^{2sj+1}\lambda} f_\gamma.$$

The damping factor depends on the scale and decreases when scale increases for $s > 0$, so that the coefficients on finer scales are damped more than coefficients on coarser scales. See Figure 4.6 for illustration.

The penalty $|\cdot|_{B^s_{1,1}(I)}$, soft shrinkage

The penalty $|\cdot|_{B^s_{1,1}(I)}$ is expressed in wavelet coefficients in the following way

$$|f|_{B^s_{1,1}(I)} \asymp \sum_{\gamma\in\Gamma} 2^{j(s-d/2)} |f_\gamma| .$$

It is clear by the results of the previous subsection, that the minimizer of the functional

$$\|u - f\|^2_{L^2(I)} + 2\lambda |u|_{B^s_{1,1}(I)}$$

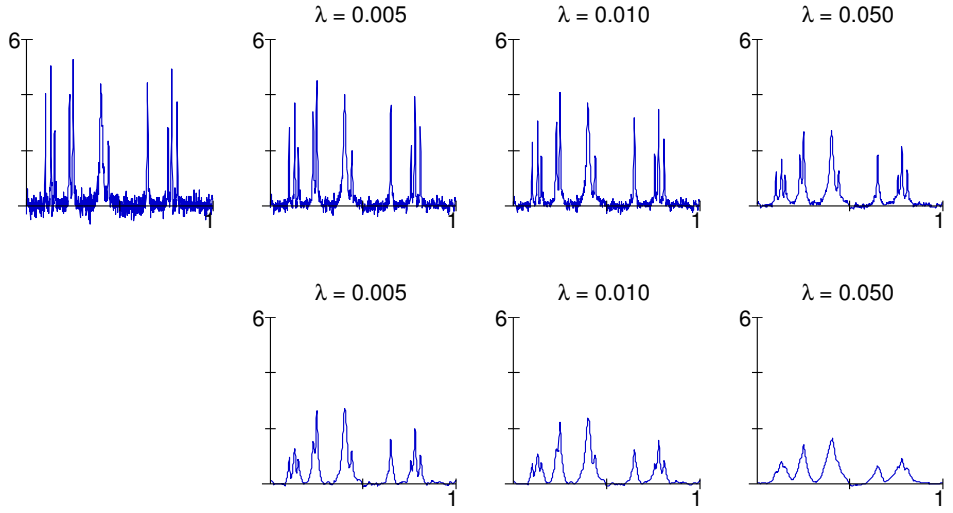

Figure 4.6: Denoising of a noisy signal with the $|\cdot|^2_{B^s_{2,2}(I)}$ penalty. Top left: Noisy signal. Upper row: $s = 1/2$, lower row: $s = 1$.

is given by

$$u_\gamma = S_{2^{j(s-d/2)}\lambda}(f_\gamma)$$

where S denotes the soft shrinkage function. For $s = d/2$ the threshold is independent of the scale and for $s > d/2$ it is increasing with scale and thus is results in a stronger smoothing effect. The effect of this shrinkage method is illustrated in Figure 4.7.

The penalty $|\cdot|^{1/2}_{B^s_{1/2,1/2}(I)}$, something between soft and hard shrinkage

To illustrate the behavior of the minimizer for $0 < p < 1$, we show, what happens for $p = 1/2$. In this case, the Besov seminorm is expressed in wavelet coefficients like

$$|f|^{1/2}_{B^s_{1/2,1/2}(I)} \asymp \sum_{\gamma\in\Gamma} 2^{j(\frac{s}{2}-\frac{3d}{4})}\,|f_\gamma|^{1/2}$$

and the minimizer of the functional $\|u-f\|^2_{L^2(I)} + 2\lambda\,|u|^{1/2}_{B^s_{1/2,1/2}(I)}$ has the coefficients

$$u_\gamma = P^{1/2}_{2^{j(\frac{s}{2}-\frac{3d}{4})}\lambda}(f_\gamma).$$

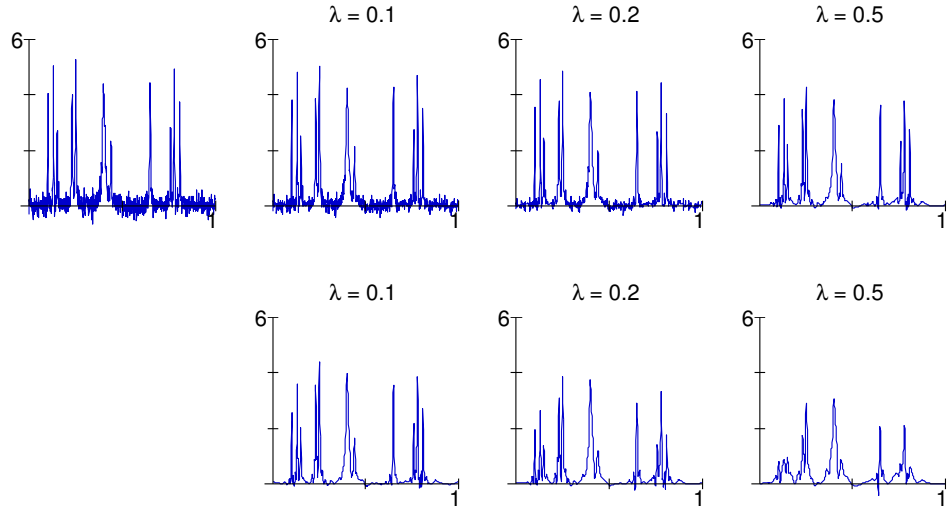

Figure 4.7: Denoising of a noisy signal with the $|\cdot|_{B^s_{1,1}(I)}$ penalty. Top left: Noisy signal. Upper row: $s = 1/2$, lower row: $s = 1$.

The function $P_\lambda^{1/2}$ is defined in Theorem 4.2. As stated in this theorem, the shrinkage function $P^{1/2}_{2^{j(\frac{s}{2}-\frac{3d}{4})}\lambda}$ has a "dead zone" for $|x| \leq \lambda_{\text{eff}}$ and the effective threshold on scale j is

$$\lambda_{\text{eff}} = \frac{3}{2} 2^{j(\frac{s}{3}-\frac{d}{2})} \lambda^{2/3}.$$

For $s = 3d/2$ the effective threshold is equal on all scales, for $s > 3d/2$ it is increasing with scale. Figure 4.8 illustrates the behavior of this shrinkage method.

The limit $p \to 0$, hard shrinkage

The case $p = 0$ can not be treated as simple as the other cases. The Besov spaces are not defined for $p = 0$. The corresponding expressions for the norms are not even quasi norms. Actually the term "space" is not accurate in this case.

But one could ask, what happens to the expressions

$$\Phi^p(f) := |f|^p_{B^s_{p,p}(I)} \asymp \sum_{\gamma \in \Gamma} 2^{sjp} 2^{j(p-2)d/2} |f_\gamma|^p$$

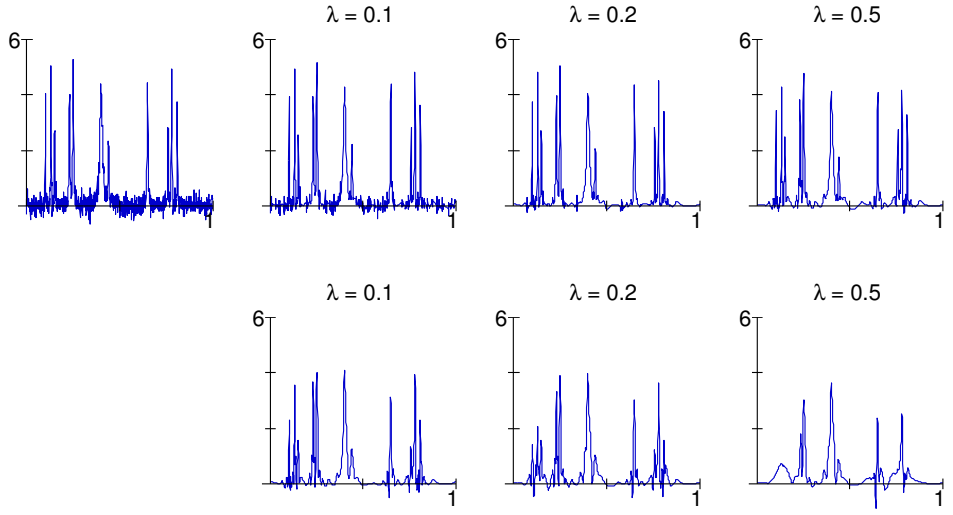

Figure 4.8: Denoising of a noisy signal with the $|\cdot|^{1/2}_{B^s_{1/2,1/2}(I)}$ penalty. Top left: Noisy signal. Upper row: $s = 3/2$, lower row: $s = 3$.

for $p \to 0$. If one assumes a coupling of s and p, one has to distinguish different cases.

First we consider, that s and p are related through $s = \left(\frac{1}{p} - \frac{1}{2}\right) d$. In this case the Besov semi norm is

$$|f|^p_{B^s_{p,p}(I)} \asymp \sum_{\gamma\in\Gamma} |f_\gamma|^p$$

and the Besov space $B^s_{p,p}(I)$ is embedded in $L^2(I)$ for $p < 2$. In fact these Besov spaces are the spaces of minimal smoothness, embedded in $L^2(I)$ [CDLL98]. Letting $p \to 0$ with s and p related as above (which means $s \to \infty$) leads to the functional

$$\Phi^0(f) := \sum_{\gamma\in\Gamma} \operatorname{sign}(|f_\gamma|) = \#\{\gamma \in \Gamma |\, f_\gamma \neq 0\}$$

which is just the number of nonzero wavelet coefficients. This penalty term leads to uniform hard shrinkage with threshold $\sqrt{2\lambda}$.

A level dependent hard shrinkage can be achieved if one couples s and p through $s = \frac{a}{p} + \left(\frac{1}{p} - \frac{1}{2}\right) d$ for $a \in \mathbb{R}$. Then the functional Φ^p converges for $p \to 0$ to

$$\tilde{\Phi}^0(f) := \sum_{\gamma\in\Gamma} 2^{aj} \operatorname{sign}(|f_\gamma|)$$

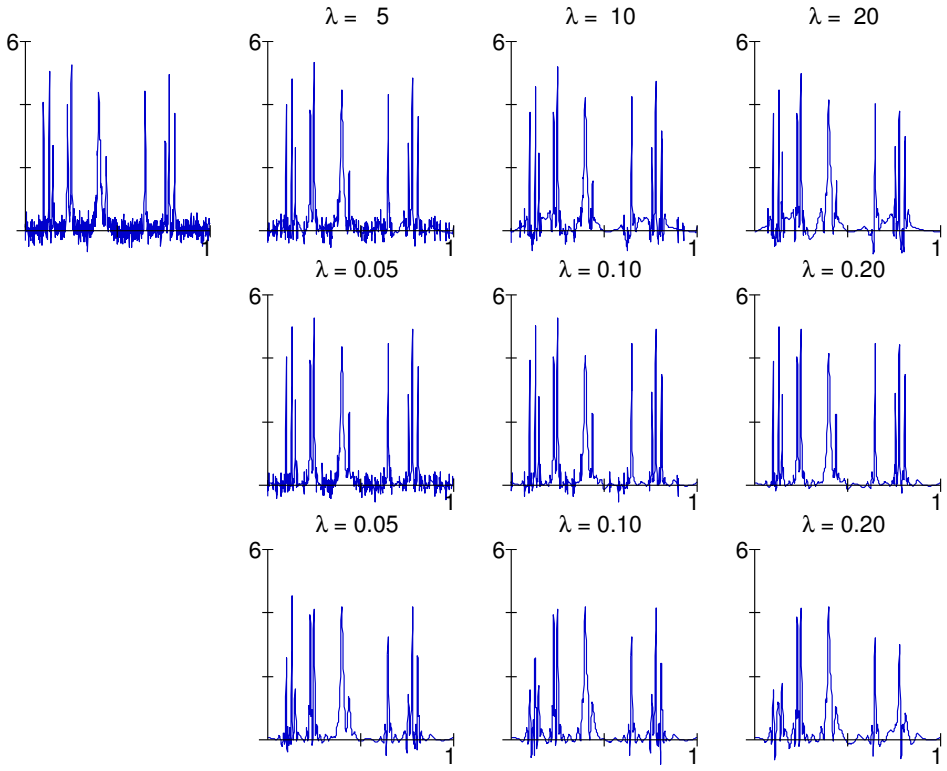

Figure 4.9: Denoising of a noisy signal with the $|\cdot|^p_{B^s_{p,p}(I)}$ penalty for $p \to 0$. Top left: Noisy signal. Upper row: $s = -1/p + (1/p - 1/2)d$, i. e. hard shrinkage where the threshold decreases with scale like 2^{-j}, middle row: $s = (1/p - 1/2)d$, i. e. hard shrinkage with uniform threshold, lower row: $s = 1/p + (1/p - 1/2)d$, i. e. hard shrinkage where the threshold increases with scale like 2^j.

and this clearly leads to the threshold $2^{aj}\sqrt{2\lambda}$ on scale j. Depending on a this is increasing or decreasing with scale.

We see, that it is not clear, what happens in the case $p \to 0$. It depends on the coupling of s and p whether we get a scale dependent or an uniform hard shrinkage. The different cases are illustrated in Figure 4.9.

4.2 The scale of penalties $|\cdot|_{B^s_{p,p}(I)}$

The aim of this section is to calculate the minimizer of the functional

$$\|u - f\|^2_{L^2(I)} + 2\lambda\, |u|_{B^s_{p,p}(I)}\,.$$

The only difference to the previous section is, that the smoothness measure omits the power p. As we will see, the resulting minimizers are quite similar to the minimizers calculated in the previous section. But the mathematical techniques which will be used are very different. A main tool will be projection onto convex sets in Hilbert spaces.

Furthermore this will provide another new understanding and new motivation of shrinkage methods. In terms of projections the shrinkage will appear as (Identity − Projection). In addition it will be possible to investigate the case $p = \infty$ which has not been treated in the previous section.

4.2.1 Minimization as a projection

Here we provide a framework for the minimization of a more general class of functionals, than we need. We are going to show, how functionals of the form

$$F(u) = \|u - f\|_H^2 + 2\lambda\Phi(u)$$

can be minimized when the functional $\Phi : H \to \overline{\mathbb{R}}$ is convex, proper, lower semi continuous and positive homogeneous and H is a real Hilbert space. We will come back to the case with the Besov penalty term in the next subsection.

The main theorem of this subsection is the following, which is a kind of duality result from convex analysis. It has been first applied in mathematical image processing in the case of total variation minimization in [Cha04].

Theorem 4.3 (Minimization by projections). *Let $f \in H$ and $F : H \to \overline{\mathbb{R}}$ be defined by*

$$F(u) := \|u - f\|_H^2 + 2\lambda\Phi(u)$$

where $\Phi : H \to \overline{\mathbb{R}}$ is convex, proper, lower semi continuous and positive homogeneous of degree one. Then the minimizer of F is

$$\tilde{u} = (\mathrm{Id} - \Pi_{\lambda C})(f)$$

where Π_C is the orthogonal projection onto the convex set

$$C := \{\, v \in H |\, \langle v |\, w \rangle \leq \Phi(w) \text{ for all } w \in H \}\,.$$

Proof. The functional F is convex, because the norm and Φ are convex. Thus the minimizers of F are characterized by

$$0 \in \partial F(u).$$

Because the mapping $u \mapsto \|u - f\|_H^2$ is differentiable, its subgradient is $\{2(u - f)\}$. With the help of the theorem of Moreau-Rockafellar (see [RW98], for example) the subgradient of F is $\partial F(u) = 2(u - f) + 2\lambda\partial\Phi(u)$. Hence the characterization of the minimizers of F looks like

$$\frac{f - u}{\lambda} \in \partial\Phi(u).$$

The inversion rule for subgradients (Theorem 2.14) says that this is equal to

$$u \in \partial\Phi^* \left(\frac{f - u}{\lambda} \right)$$

and further transformation gives

$$u = f - \lambda \left(\mathrm{Id} + \frac{1}{\lambda}\partial\Phi^* \right)^{-1} \left(\frac{f}{\lambda} \right).$$

At the first glance this looks quite difficult. But the dual function of Φ has in fact a quite simple structure: Because Φ is positive homogeneous of degree one its dual function is the indicator function of a certain set (see Example 2.12). It holds

$$\Phi^* = \delta_C$$

where the set C is defined by $C := \{ v \in H | \, \langle v | \, w \rangle \leq \Phi(w) \text{ for all } w \in H \}$. Especially we have $\lambda\partial\Phi^* = \partial\Phi^*$ or all $\lambda > 0$. Now we show that $(\mathrm{Id} + \partial\Phi^*)^{-1}$ is the projection onto C. Let us assume, that

$$w = (\mathrm{Id} + \partial\Phi^*)^{-1} (v).$$

This means, that $w - v + \partial\Phi^*(w) \ni 0$ and hence w is a minimizer of

$$\|v - w\|_H^2 + 2\Phi^*(w).$$

Because Φ^* is the indicator function of C this is exactly the condition which characterizes the orthogonal projection onto C.

The remark that for a projection Π_C, $f \in H$ and $\lambda > 0$ it holds

$$\lambda \, \Pi_C(f/\lambda) = \Pi_{\lambda C}(f)$$

completes the proof. □

4.2.2 The Besov penalties

To apply the results of the previous subsection to the functionals

$$\|u-f\|^2_{L^2(I)} + 2\lambda\,|u|_{B^s_{p,p}(I)}$$

we express this functional through the wavelet coefficients of f and u.

Using the norm equivalences stated in Subsection 2.2.4 we can work in the sequence space $\ell^2(\Gamma)$ instead of function spaces. We obtain the functional $F : \ell^2(\Gamma) \to \overline{\mathbb{R}}$ defined by

$$F((u_\gamma)) = \sum_{\gamma\in\Gamma}(u_\gamma - f_\gamma)^2 + 2\lambda\Phi((u_\gamma))$$

with the positive homogeneous functional

$$\Phi((u_\gamma)) = \left(\sum_{\gamma\in\Gamma} 2^{sjp}2^{j(p-2)d/2}\,|u_\gamma|^p\right)^{\frac{1}{p}}.$$

Theorem 4.3 states, that this minimization is solved by a projection. To show, how the projections look like, we distinguish three cases: $1 < p < \infty$, $p = 1$ and $p = \infty$. In this framework it is not possible to treat the case $p < 1$ because the functional Φ is no more convex and the duality between positive homogeneous functions and convex sets, which only holds for convex functions, is essential.

In the next theorem it is shown, how the projection looks like for $1 < p < \infty$.

Theorem 4.4 (Besov penalties and projections, $1 < p < \infty$). *Let $(f_\gamma) \in \ell^2(\Gamma)$ and $1 < p < \infty$. The functional*

$$F((u_\gamma)) = \sum_{\gamma\in\Gamma}(u_\gamma - f_\gamma)^2 + 2\lambda\Phi((u_\gamma))$$

with the positive homogeneous functional

$$\Phi((u_\gamma)) = \left(\sum_{\gamma\in\Gamma} 2^{sjp}2^{j(p-2)d/2}\,|u_\gamma|^p\right)^{\frac{1}{p}}$$

is minimized by $(\mathrm{Id} - \Pi_{\lambda C})(f)$ and the set C is

$$C = \left\{ v \in \ell^2(\Gamma) \;\middle|\; \left(\sum_{\gamma\in\Gamma} 2^{-jsp^*}2^{j(p^*-2)d/2}\,|v_\gamma|^{p^*}\right)^{\frac{1}{p^*}} \le 1 \right\}$$

where p^* is the dual exponent to p defined by $1/p + 1/p^* = 1$.

Before we prove this theorem we provide the following lemma which calculates dual functions for weighted $\ell^p(\Gamma)$ norms in $\ell^2(\Gamma)$.

Lemma 4.5. *Let $1 < p < \infty$, (a_γ) be a sequence with $a_\gamma > 0$ for every γ and let $\Phi : \ell^2(\Gamma) \to \overline{\mathbb{R}}$ be defined by*

$$\Phi(u) = \left(\sum_{\gamma \in \Gamma} a_\gamma \left| u_\gamma \right|^p \right)^{\frac{1}{p}}.$$

Then $\Phi^* = \delta_C$ and the set C is

$$C = \left\{ v \in \ell^2(\Gamma) \;\middle|\; \left(\sum_{\gamma \in \Gamma} a_\gamma^{1-p^*} \left| v_\gamma \right|^{p^*} \right)^{\frac{1}{p^*}} \leq 1 \right\}$$

Proof. We are going to show, that $(\delta_C)^* = \Phi$. This gives the assertion because Φ is convex (see Remark 2.11). To do so we have to show that

$$\Phi(u) = \sup_v \Big(\langle u | v \rangle - \delta_C(v) \Big) = \sup_{v \in C} \langle u | v \rangle .$$

By Hölder's inequality we have

$$\begin{aligned} \sum_{\gamma \in \Gamma} u_\gamma v_\gamma &\leq \sum_{\gamma \in \Gamma} \left| a_\gamma^{1/p} u_\gamma \right| \left| a_\gamma^{-1/p} v_\gamma \right| \\ &\leq \left(\sum_{\gamma \in \Gamma} a_\gamma \left| u_\gamma \right|^p \right)^{1/p} \left(\sum_{\gamma \in \Gamma} a_\gamma^{-p^*/p} \left| v_\gamma \right|^{p^*} \right)^{1/p^*} \end{aligned}$$

Using $-p^*/p = 1 - p^*$ and taking suprema over $v \in C$ on both sides yields

$$\sup_{v \in C} \langle u | v \rangle \leq \Phi(u).$$

It remains to find elements $u \in \ell^2(\Gamma)$ and $v \in C$ for which the above inequality is sharp. We take $v_\gamma = c a_\gamma \operatorname{sign}(u_\gamma) \left| u_\gamma \right|^{p-1}$, where c is a constant which guarantees

$$\sum_{\gamma \in \Gamma} a_\gamma^{1-p^*} \left| v_\gamma \right|^{p^*} = 1$$

and especially $v \in C$ (actually the constant is $c = \left(\sum_{\gamma\in\Gamma} a_\gamma \left|u_\gamma\right|^p\right)^{-1/p^*}$). With this choice we get

$$\begin{aligned}\sum_{\gamma\in\Gamma} v_\gamma u_\gamma &= c \sum_{\gamma\in\Gamma} a_\gamma \left|u_\gamma\right|^p \\ &= c \left(\sum_{\gamma\in\Gamma} a_\gamma \left|u_\gamma\right|^p\right)^{1/p} \left(\sum_{\gamma\in\Gamma} a_\gamma \left|u_\gamma\right|^p\right)^{1-1/p}\end{aligned}$$

and because of $a_\gamma \left|u_\gamma\right|^p = c^{-p^*} a_\gamma^{1-p^*} \left|v_\gamma\right|^{p^*}$ we see

$$\sum_{\gamma\in\Gamma} u_\gamma v_\gamma = \left(\sum_{\gamma\in\Gamma} a_\gamma \left|u_\gamma\right|^p\right)^{1/p} \left(\sum_{\gamma\in\Gamma} a_\gamma^{1-p^*} \left|v_\gamma\right|^{p^*}\right)^{1/p^*}.$$

This shows

$$\langle u | v \rangle = \Phi(u) \left(\sum_{\gamma\in\Gamma} a_\gamma^{1-p^*} \left|v\right|^{p^*}\right)^{1/p^*} = \Phi(u)$$

and the proof is complete. □

Now the proof of Theorem 4.4 is almost done.

Proof of Theorem 4.4. All which remains to show is that the set C has the stated form. We apply the above lemma for the weight $a_\gamma = 2^{sjp} 2^{j(p-2)d/2}$ and get that

$$C = \left\{ v \in \ell^2(\Gamma) \;\middle|\; \left(\sum_{\gamma\in\Gamma} \left(2^{jsp} 2^{j(p-2)d/2}\right)^{1-p^*} \left|v_\gamma\right|^{p^*}\right)^{\frac{1}{p^*}} \le 1 \right\}.$$

The identities $p(1-p^*) = -p^*$ and $(p-2)(1-p^*) = p^* - 2$ show

$$\left(2^{jsp} 2^{j(p-2)d/2}\right)^{1-p^*} = 2^{-jsp^*} 2^{j(p^*-2)d/2}$$

which completes the proof. □

The cases $p = 1$ and $p = \infty$ has to be treated separately and the arguments are simpler. For $p = 1$ the Besov norm looks like

$$|f|_{B^s_{1,1}(I)} \asymp \sum_{\gamma\in\Gamma} 2^{j(s-d/2)} \left|f_\gamma\right|$$

and we have the following result.

Theorem 4.6 (Besov penalties and projections, $p = 1$). *Let $(f_\gamma) \in \ell^2(\Gamma)$. The functional*

$$F((u_\gamma)) = \sum_{\gamma\in\Gamma}(u_\gamma - f_\gamma)^2 + 2\lambda\Phi((u_\gamma))$$

with the positive homogeneous functional $\Phi((u_\gamma)) = \sum_{\gamma\in\Gamma} 2^{j(s-d/2)}|u_\gamma|$ is minimized by $(\mathrm{Id} - \Pi_{\lambda C})(f)$ and the set C is

$$C = \left\{ v \in \ell^2(\Gamma) \,\middle|\, \sup_{\gamma\in\Gamma} 2^{-j(s-d/2)}|v_\gamma| \le 1 \right\}$$

Proof. The proof that $\langle u|\, v\rangle \le \Phi(u)$ holds for v which fulfills

$$\sup_{\gamma\in\Gamma} 2^{-j(s-d/2)}|v_\gamma| \le 1$$

is similar to the proof of the above lemma. If we conversely assume, that $\sup_{\gamma\in\Gamma} 2^{-j(s-d/2)}|v_\gamma| > 1$ then we have at least one index γ_0 for which

$$2^{-j_0(s-d/2)}|v_{\gamma_0}| > 1.$$

We choose

$$u_\gamma = \begin{cases} \operatorname{sign} v_{\gamma_0} & \text{for } \gamma = \gamma_0 \\ 0 & \text{else} \end{cases}$$

which yields

$$\langle u|\, v\rangle = |v_{\gamma_0}| > 2^{j_0(s-d/2)} = \Phi(u)$$

and this shows $v \notin C$. □

Finally we treat the case $p = \infty$. In this case the Besov semi norm is equivalent to

$$|f|_{B^s_{\infty,\infty}(I)} \asymp \sup_{\gamma\in\Gamma} 2^{j(s+d/2)}|f_\gamma|$$

and the following result holds.

Theorem 4.7 (Besov penalties and projections, $p = \infty$). *Let $(f_\gamma) \in \ell^2(\Gamma)$. The functional*

$$F((u_\gamma)) = \sum_{\gamma\in\Gamma}(u_\gamma - f_\gamma)^2 + 2\lambda\Phi((u_\gamma))$$

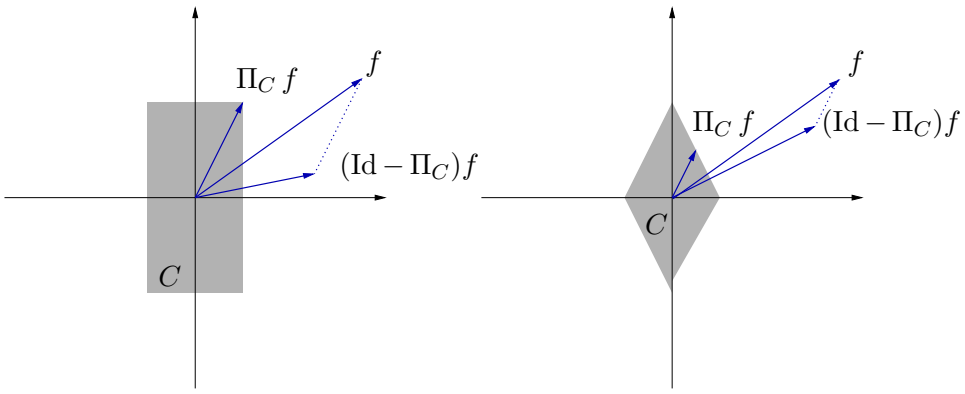

Figure 4.10: Illustration of the effect of the operator $(\mathrm{Id}-\Pi_C)$. On the left the set C is a ball corresponding to a weighted ∞-norm and on the right it is a weighted 1-ball. The operator $(\mathrm{Id}-\Pi_C)$ is in both cases a shrinkage.

with the positive homogeneous functional $\Phi((u_\gamma)) = \sup_{\gamma\in\Gamma} 2^{j(s+d/2)}\,|u_\gamma|$ *is minimized by* $(\mathrm{Id}-\Pi_{\lambda C})(f)$ *and the set* C *is*

$$C = \left\{ v \in \ell^2(\Gamma) \,\middle|\, \sum_{\gamma\in\Gamma} 2^{-j(s+d/2)}\,|v_\gamma| \le 1 \right\}$$

Proof. To show, that $\sum_{\gamma\in\Gamma} 2^{-j(s+d/2)}\,|v_\gamma| \le 1$ implies $\langle u|\, v\rangle \le \Phi(u)$ is similar to the case $p = 1$.

To find elements u and v for which the upper inequality is sharp, we note that for u satisfying $\Phi(u) < \infty$ and every $\varepsilon > 0$ we can find an index γ_0 such that $2^{j_0(s+d/2)}\,|u_{\gamma_0}| \ge \sup_{\gamma\in\Gamma}\left(2^{j(s+d/2)}\,|u_\gamma|\right) - \varepsilon$. We define v by

$$v_\gamma = \begin{cases} 2^{j_0(s+d/2)}\,\mathrm{sign}(u_{\gamma_0}) & \text{for } \gamma = \gamma_0 \\ 0 & \text{else.} \end{cases}$$

This yields

$$\langle u|\, v\rangle = 2^{j_0(s+d/2)}\,|u_{\gamma_0}| \ge \sup_{\gamma\in\Gamma}\left(2^{j(s+d/2)}\,|u_\gamma|\right) - \varepsilon = \Phi(u) - \varepsilon.$$

Letting $\varepsilon \to 0$ gives the assertion. □

4.2.3 Special cases and illustrations

The preceding three theorems showed, that the stated variational problems are solved by an orthogonal projection in the space of wavelet coefficients. Comprisingly one can say, that the minimization problem

$$\|u - f\|_2^2 + 2\lambda \, |u|_{B^s_{p,p}(I)}$$

is solved by subtracting a $B^{-s}_{p^*,p^*}(I)$-part of size λ from the original image or signal f (compare Subsection 2.2.3 and Theorems 4.4, 4.6 and 4.7).

This fact has a few nice interpretations: We consider the space $B^1_{1,1}(I)$ in two dimensions. This space is known to be very close to the space of functions of bounded variation $BV(I)$ [CDDD03]. The space $BV(I)$ plays a very important role in denoising. In the paper [ROF92] Rudin, Osher and Fatemi proposed the $BV(I)$ semi norm as penalty term for our variational functional. This leads to so called total variation methods which has proven to be very efficient (see [SC96, Ves01, SC03, Kee03], for example, for both theoretical results and practical applications). The exact numerical implementation of total variation methods is a hard task and several methods have been proposed [DV97, CL97, HK04, Cha04]. The space $B^1_{1,1}(I)$ is interesting because it is close to the space $BV(I)$ and furthermore it leads to a very simple denoising algorithm compared to the BV denoising. The dual space of $B^1_{1,1}(I)$ is $B^{-1}_{\infty,\infty}(I)$ and this space is used to model textures [Mey01, AA03]. The semi norm in $B^{-1}_{\infty,\infty}(I)$ is suitable to measure oscillations. Our result states, that the minimization with a $B^1_{1,1}(I)$ penalty term is just the subtraction of an oscillating part of the size λ.

For other penalty terms there are similar interpretations. For example one can use the semi norm $|\cdot|_{H^{d/2}(I)}$ as penalty term. Then the solution of the corresponding minimization problem is the original signal f minus a $H^{-d/2}(I)$-component of f of size λ. The Sobolev space $H^{-d/2}(I)$ is the smallest Sobolev space which contains white noise and one could say, that this minimization wipes out a white noise component.

The important role that dual norms play in denoising was recognized and investigated by several people and gave rise to new promising denoising methods [OSV03, MH04, AC04, OOS04]

In the following we will treat the three special cases $p = 1$, $p = 2$ and $p = \infty$ separately. For $p = 1$ we will rediscover the soft wavelet shrinkage. The case $p = 2$ leads to a level dependent damping and $p = \infty$ yields in a "clipping", i. e. the large wavelet coefficients are reduced to a certain value.

The penalty $|\cdot|_{B^s_{1,1}(I)}$

Theorem 4.6 states, that the projection, we have to compute is the projection onto the weighted ∞-ball given by

$$\lambda C = \left\{ v \in \ell^2(\Gamma) \,\middle|\, \sup_{\gamma\in\Gamma} 2^{-j(s-d/2)} |v_\gamma| \le \lambda \right\}.$$

One easily sees, that this projection is performed by

$$\Pi_{\lambda C}((f_\gamma)) = (C_{2^{j(s-d/2)}\lambda}(f_\gamma))$$

where the clipping function

$$C_\mu(x) = \begin{cases} \mu & , x \ge \mu \\ x & , |x| < \mu \\ -\mu & , x \le -\mu \end{cases}$$

is applied component wise to the coefficients f_γ. Hence the minimizer of

$$\|u - f\|^2_{L^2(I)} + 2\lambda |u|_{B^s_{1,1}(I)}$$

turns out to be

$$\begin{aligned} u &= (\mathrm{Id} - \Pi_{\lambda C})(f) \\ &= \langle f | 1\rangle + \sum_{\gamma\in\Gamma} (f_\gamma - C_{2^{j(s-d/2)}\lambda}(f_\gamma))\, \psi_\gamma \\ &= \langle f | 1\rangle + \sum_{\gamma\in\Gamma} (S_{2^{j(s-d/2)}\lambda}(f_\gamma))\, \psi_\gamma. \end{aligned}$$

For $s = d/2$ this is the uniform soft shrinkage again. This time it occurs as "identity minus the projection onto the ∞-ball". This is the third place, where the soft shrinkage function appears in a very different context from denoising: As the function

$$S_\lambda(x) = x - \Pi_{[-\lambda,\lambda]}(x).$$

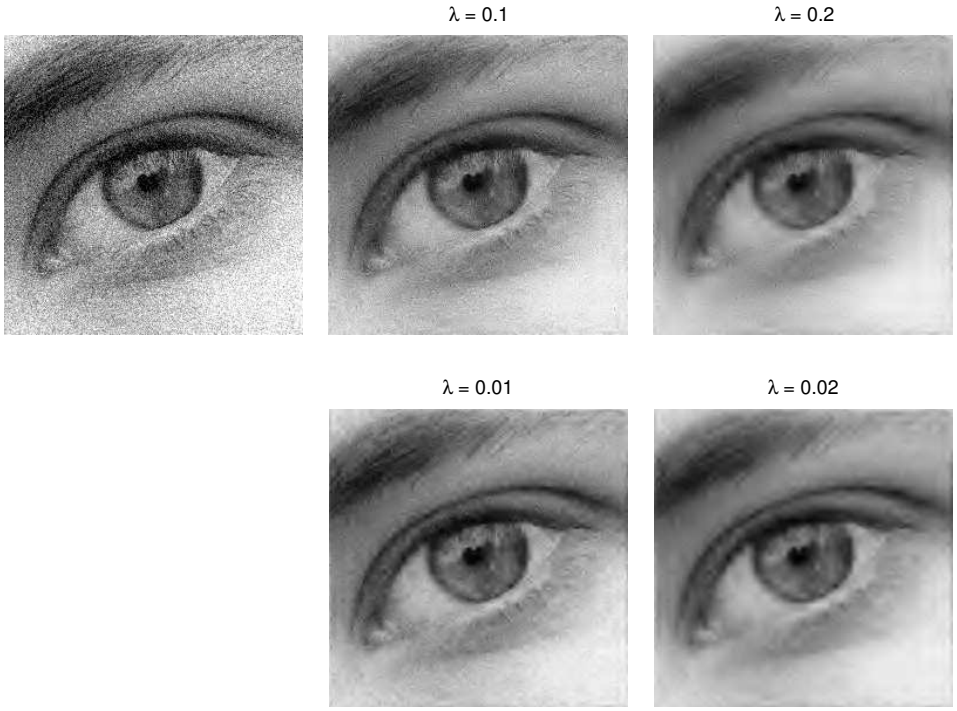

Figure 4.11: Denoising with the Besov penalty $B^s_{1,1}(I)$ which corresponds to a non linear scale dependent soft shrinkage of the wavelet coefficients. On the left is the original noisy image. Upper row: Denoised with the $B^1_{1,1}(I)$ penalty. Lower row: Denoised with the $B^2_{1,1}(I)$ penalty.

The penalty $|\cdot|_{B^s_{2,2}(I)}$, Sobolev spaces

Another interesting case is the scale $|\cdot|_{B^s_{2,2}(I)}$ of penalties. These Besov spaces coincide with the Sobolev spaces $H^s(I)$. The projection, we have to calculate is the orthogonal projection onto the set

$$\lambda C = \left\{ v \in \ell^2(\Gamma) \,\middle|\, \sum_{\gamma\in\Gamma} 2^{-2js} |v_\gamma|^2 \le \lambda^2 \right\}.$$

This projection is characterized by the constrained minimization problem

$$\text{minimize } \sum_{\gamma\in\Gamma} (v_\gamma - f_\gamma)^2 \text{ subject to } \sum_{\gamma\in\Gamma} 2^{-2js} |v_\gamma|^2 \le \lambda^2.$$

The solution of this problem can be obtained with the help of Lagrange multipliers. This leads to the solution of the equations

$$(v_\gamma - f_\gamma) + \mu 2^{-2js} v_\gamma = 0$$

where $\mu > 0$ is the Lagrange multiplier. Hence the solution is

$$\bar{v}_\gamma = \frac{f_\gamma}{1 + \mu 2^{-2js}}.$$

To show, that such a Lagrange multiplier exists, we plug this minimizer into the constraint:

$$\lambda^2 = \sum_{\gamma \in \Gamma} \frac{2^{-2js}}{(1 + \mu 2^{-2js})^2} |f_\gamma|^2 .$$

The right hand side is monotonically decreasing and continuous in μ and decreases from $|f|^2_{H^{-s}(I)}$ to zero if μ increases from 0 to ∞. This shows, that there is a Lagrange multiplier μ such that $(\bar{v}_\gamma)$ is the projection if $|f|_{H^{-s}(I)} \geq \lambda$.

The minimizer of

$$\|u - f\|^2_{L^2(I)} + 2\lambda |u|_{H^s(I)}$$

hence is

$$\begin{aligned} u &= (\mathrm{Id} - \Pi_{\lambda C})(f) \\ &= \langle f | 1 \rangle + \sum_{\gamma \in \Gamma} \left(f_\gamma - \frac{f_\gamma}{1 + \mu 2^{-2js}} \right) \psi_\gamma \\ &= \langle f | 1 \rangle + \sum_{\gamma \in \Gamma} \left(\frac{\mu 2^{-2js}}{1 + \mu 2^{-2js}} f_\gamma \right) \psi_\gamma \\ &= \langle f | 1 \rangle + \sum_{\gamma \in \Gamma} \left(\frac{1}{1 + 2^{2js+1} \frac{1}{2\mu}} f_\gamma \right) \psi_\gamma . \end{aligned}$$

Compare this result to the results from Subsection 4.1.3 where the penalty term $|\cdot|^2_{H^2(I)}$ yielded the damping factor $\frac{1}{1+2^{2js+1}\lambda}$ on scale j, i. e. the shrinkage is exactly the same and the Lagrange multiplier μ and the parameter λ for the penalty $|\cdot|^2_{H^s(I)}$ are related through $\lambda = \frac{1}{2\mu}$.

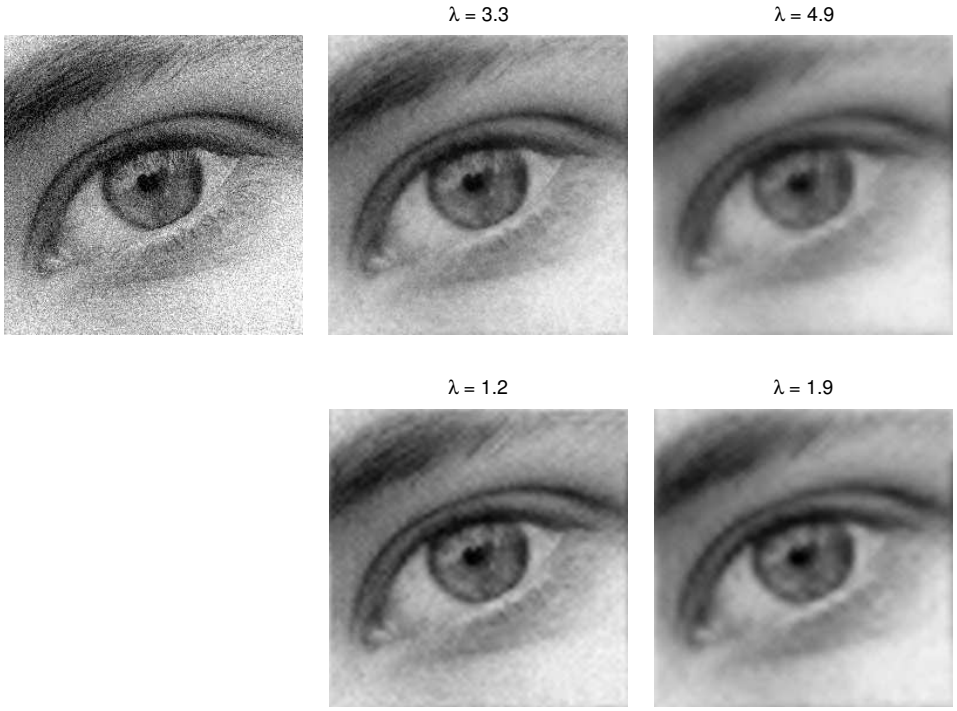

Figure 4.12: Denoising with a Sobolev penalty which corresponds to a linear scale dependent damping of the wavelet coefficients. On the left is the original noisy image. Upper row: Denoised with the $H^1(I)$ penalty. Lower row: Denoised with the $H^2(I)$ penalty.

The penalty $|\cdot|_{B^s_{\infty,\infty}(I)}$, Hölder spaces

For $p = \infty$ the argumentation is almost similar to the case $p = 2$. The projection we have to concern is the projection onto a ball corresponding to a weighted 1-norm, namely onto

$$\lambda C = \left\{ v \in \ell^2(\Gamma) \,\middle|\, \sum_{\gamma\in\Gamma} 2^{-j(s+d/2)} |v_\gamma| \le \lambda \right\}.$$

The method of Lagrange multipliers leads to the equations

$$2(v_\gamma - f_\gamma) + \mu 2^{-j(s+d/2)} \operatorname{sign}(v_\gamma) \ni 0$$

with the Lagrange multiplier $\mu > 0$. The operator sign is the set valued subdifferential of the absolute value. The solution is

$$\bar{v}_\gamma = S_{\frac{\mu}{2}2^{-j(s+d/2)}}(f_\gamma).$$

The Lagrange multiplier is determined by the constraint condition

$$\lambda = \sum_{\gamma\in\Gamma} 2^{-j(s+d/2)} S_{\frac{\mu}{2}2^{-j(s+d/2)}}(|f_\gamma|).$$

To show the existence of the Lagrange multiplier, we denote the right hand side of the above equation by $h(\mu)$ and note again, that h is decreasing from $|f|_{B^{-s}_{1,1}(I)}$ to zero if μ increases from zero to ∞. It remains to show, that h depends continuously on μ. We estimate the difference

$$\begin{aligned}
h(\mu+\delta)-h(\mu) &= \sum_{|f_\gamma|\geq\frac{\mu+\delta}{2}2^{-j(s+d/2)}} 2^{-j(s+d/2)}(|f_\gamma| - \frac{\mu+\delta}{2}2^{-j(s+d/2)}) \\
&\quad - \sum_{|f_\gamma|\geq\frac{\mu}{2}2^{-j(s+d/2)}} 2^{-j(s+d/2)}(|f_\gamma| - \frac{\mu}{2}2^{-j(s+d/2)}) \\
&\leq \sum_{|f_\gamma|\geq\frac{\mu+\delta}{2}2^{-j(2s+d)}} 2^{-j(s+d/2)}(-\frac{\delta}{2}2^{-j(s+d/2)}) \\
&\leq -\frac{\delta}{2}\sum_{j=0}^{\infty}\sum_{k\in\Gamma_j}\sum_{i=1}^{2^d-1} 2^{-j(2s+d)}.
\end{aligned}$$

We use, that the set Γ_j has 2^{jd} elements (compare Subsection 1.2.1) and end up with

$$h(\mu+\delta)-h(\mu) \leq -\frac{\delta}{2}(2^d-1)\sum_{j=0}^{\infty} 2^{-js}.$$

Because the series on the right hand side converges for $s>0$, this shows, that h is continuous (even Lipschitz continuous) if $s>0$. Thus, the existence of a Lagrange multiplier is proved for $s>0$.

The minimizer of

$$\|u-f\|^2_{L^2(I)} + 2\lambda\,|u|_{B^s_{\infty,\infty}(I)}$$

hence is

$$\begin{aligned}
u &= (\mathrm{Id}-\Pi_{\lambda C})(f) \\
&= \langle f|\,1\rangle + \sum_{\gamma\in\Gamma}\Big(f_\gamma - S_{\frac{\mu}{2}2^{-j(s+d/2)}}(f_\gamma)\Big)\,\psi_\gamma \\
&= \langle f|\,1\rangle + \sum_{\gamma\in\Gamma}\Big(C_{\frac{\mu}{2}2^{-j(s+d/2)}}(f_\gamma)\Big)\,\psi_\gamma,
\end{aligned}$$

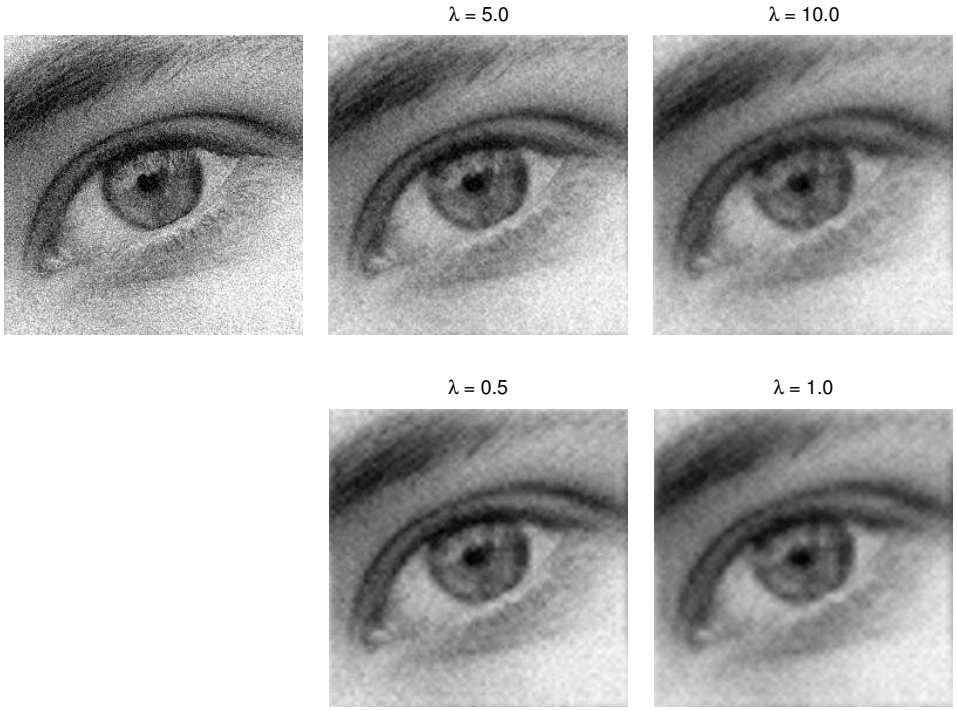

Figure 4.13: Denoising with a kind of Hölder penalty which corresponds to a nonlinear scale dependent clipping of the wavelet coefficients. On the left is the original noisy image. Upper row: Denoised with the $B^1_{\infty,\infty}(I)$ penalty. Lower row: Denoised with the $B^2_{\infty,\infty}(I)$ penalty.

where C_λ is the clipping function. The smoothing effect of this method is illustrated in Figure 4.13.

4.3 Comparison of the different shrinkage methods

In the previous two sections we showed illustrations for the different shrinkage methods separately. The focus was on the illustration of the methods and not on a comparison.

The last section of this chapter will present a comparison of all the shrinkage methods from this chapter. We use the example signal `bumps` and the image `eye` together with the methods from Section 4.1 and Section 4.2 respectively. To make the shrinkage methods comparable, we choose the

signal-to-noise-ratio (SNR) as a quality measure for denoising. The SNR is widely used to measure the performance of denoising and compresseion. The SNR of an original (assumed to be noise free) signal $\bar{f}$ and a noisy signal f is given by

$$\mathrm{SNR}(\bar{f}, f) = 10 \log_{10} \left(\frac{\|\bar{f}\|_2^2}{\|\bar{f} - f\|_2^2} \right).$$

We optimized the methods depending on the shrinkage parameter λ with respect to the SNR of the original signal (resp. image) and the denoised version.

For the comparison of the pointwise shrinkage methods from Section 4.1 we used the `sym4` wavelet. We start with a noisy `bumps`-signal f with a signal-noise-ratio $\mathrm{SNR}(\bar{f}, f) = 12.5$. The optimized results are shown in Figure 4.14 for eight different cases of shrinkage functions which have been investigated in Subsection 4.1.3 and are illustrated in the Figures 4.6 to 4.9: For each value of $p = 0, 1/2, 1, 2$ two cases are assumed. Namely these cases are:

$p = 0$:	$a = 0$,	$a = 1/2$,	hard shrinkage
$p = 1/2$:	$s = 3/2$,	$s = 3$,	between hard and soft shrinkage
$p = 1$:	$s = 1/2$,	$s = 1$,	soft shrinkage
$p = 2$:	$s = 1/2$,	$s = 1$,	damping.

For $p = 0, 1/2, 1$ the first case of a resp. s results in a scale independent shrinkage and the other one in a shrinkage where the threshold increases with scale.

The case of linear damping ($p = 2$) leads to the worst SNR. This is due to the strong deformation of the peaks. The soft shrinkage ($p = 1$) shows this effect too, but the noise reduction is much better. For hard shrinkage ($p = 0$) the peaks are preserved very good, but there are strong artifacts due to the discontinuity of the shrinkage function. The case $p = 1/2$ looks almost similar to $p = 0$ but leads to larger values of the SNR. This points to a better balancing between the deformation of the peaks and denoising that for soft and hard shrinkage.

The methods from Section 4.2 are compared using the image `eye` and the `coif4` wavelet. We add Gaussian white noise to the image which results in an SNR of 15.7.

Three cases of p and two cases of s are assumed: $p = 1, 2, \infty$, $s = 1, 2$. Thus we consider the spaces $B^1_{1,1}(I)$, $B^2_{1,1}(I)$, $H^1(I)$, $H^2(I)$, $C^1(I)$ and $C^2(I)$. The optimal denoising results with respect to the SNR are shown in Figure 4.15.

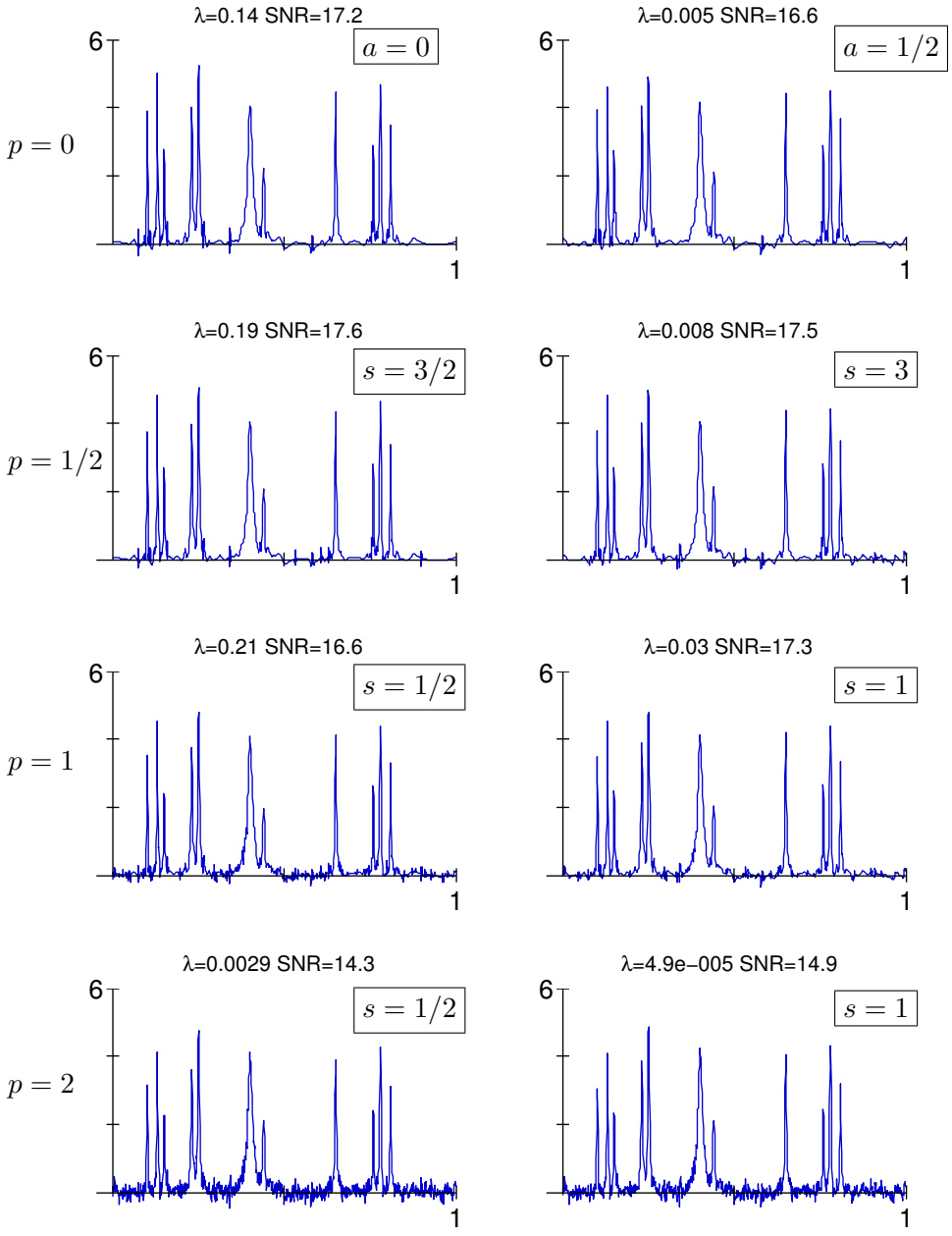

Figure 4.14: Comparison of the shrinkage methods from Section 4.1.

Though the optimal results of the different shrinkage methods are almost equal in terms of the SNR their visual impression is very different. The minimizers with respect to the $B^1_{1,1}(I)$ and the $B^2_{1,1}(I)$ penalty show a good denoising of homogeneous regions and well edge preservation. The disadvantage of these denoising methods is that more or less strong artifacts in shape of the wavelet appear. The penalties $H^1(I)$ and $H^2(I)$ lead to more blurring of the edges and a worse denoising in homogeneous regions but some small details, like the eye lashes, are preserved better. The case of Hölder spaces is most unpleasant to the eye. The edges are blurred and the noise is not removed but only smoother than before. The only good thing about this denoising is the treatment of texture like parts. As one notices, the eyebrows look most natural for these penalties.

These observations agree with the properties of the minimizers, which one would expect from the theory: The $B^s_{1,1}(I)$ spaces are close to the space of functions of bounded variation $BV(I)$ which is an appropriate space for functions with discontinuities along lines. Thus, the minimizer, which is in fact a soft shrinkage of the wavelet coefficients, shows a good preservation of edges. Some small details like the eyelashes are disappeared in the minimizer and the same behavior is known for BV minimizers [SC96, BWWS03]. The artifacts, which does not appear for BV minimizers, are due to the fact that the Besov spaces are just close to BV and not equivalent. The Sobolev spaces $H^s(I)$ and the Hölder spaces $C^s(I)$ does not contain discontinuous functions for $s = 1, 2$ and though the edges are blurred. The Sobolev spaces lead to a kind of "global shrinkage" which corresponds to their close relation to Fourier expansions. The Hölder spaces are appropriate to measure small oscillations like textures because the Hölder norm pays only attention to the largest magnitude of the derivative.

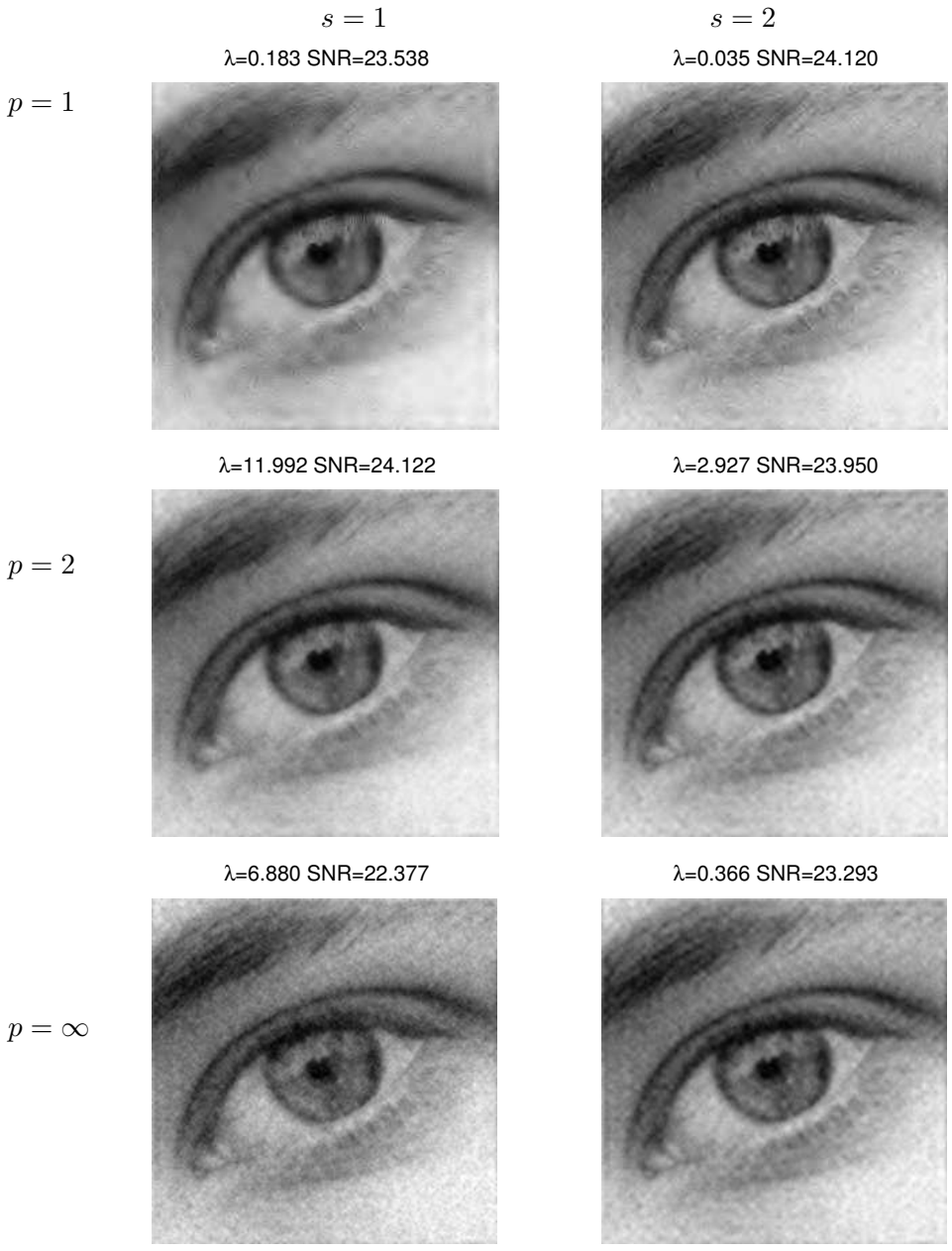

Figure 4.15: Comparison of the shrinkage methods from Section 4.2.

CHAPTER 5

Wavelet shrinkage and Bayesian estimation

This chapter will briefly present results about Bayesian denoising methods and wavelet shrinkage. It is not meant to be a detailed introduction to Bayesian statistics. We will only present the facts which are necessary to establish the equivalences between Bayesian denoising and wavelet shrinkage.

The application of Bayesian methods to the denoising problem with wavelets started with [SA96] and other references on this field are [ML99, FN01, cS02].

In this chapter we consider the following model for our noise: We have a real image f and an observed image g which is the real image corrupted by additive white noise:

$$g = f + \varepsilon$$

where ε is white noise. After a discrete wavelet transform we have

$$(g_\gamma) = (f_\gamma) + (\varepsilon_\gamma).$$

Because the wavelet transform is an orthogonal transformation, the coefficients (ε_γ) are again white noise, i. e. they are Gaussian distributed.

The goal is, to estimate the coefficients f_γ from the knowledge of g_γ and some assumptions on the distribution of the noise and the coefficients.

Because we consider the noise to be independent, we estimate the coefficient f_γ only from the knowledge of g_γ. To simplify notation, we write y for a known coefficient of g and x for the coefficient of f we want to estimate. We assume, that the distributions of the wavelet coefficients x and the noise $\varepsilon = y - x$ is known.

The "maximum a posteriori" estimator for x is

$$\hat{x}(y) = \operatorname*{argmax}_x p(x|y)$$

where $p(x|y)$ is the probability that x is the true value under the assumption, that y is observed. Thus the maximum a posteriori estimator is the value $\hat{x}$ which is the most likely from the knowledge that y has been observed.

Remark 5.1. *Note the slight but important difference to the maximum likelihood estimator which is*

$$\hat{x}(y) = \operatorname*{argmax}_x p(y|x).$$

The maximum likelihood estimator for x is the value which makes the observation y most probable.

To calculate the maximum a posteriori estimator one uses Bayes rule for the conditional probability and gets

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

Hence the maximum a posteriori estimator is

$$\hat{x}(y) = \operatorname*{argmax}_x p(y|x)p(x)$$

or, after taking logarithms,

$$\hat{x}(y) = \operatorname*{argmax}_x \log(p(y|x)) + \log(p(x)).$$

Special assumptions on the distributions of the coefficients x and on the conditional distribution $y|x$ lead to explicit formulas for the maximum a posteriori estimator.

Because we have an additive model of the noise, the probability distribution of $y|x$ is the distribution of the noise $\epsilon = y - x$ which is Gaussian:

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2\sigma^2}}.$$

It is a well known observation, that the wavelet coefficients of natural images have a non-Gaussian distribution which is often modeled as Laplacian

$$p(x) = \frac{1}{\eta\sqrt{2}} e^{-\frac{\sqrt{2}|x|}{\eta}}.$$

If we plug this into the maximum a posteriori estimator for x we end up with

$$\hat{x}(y) = \operatorname*{argmin}_x (y-x)^2 + \frac{2\sqrt{2}\sigma^2}{\eta} |x| .$$

From Section 4.1 we know, that this minimizer is just the result of soft shrinkage

$$\hat{x}(y) = S_{\frac{\sqrt{2}\sigma}{\eta}}(y).$$

This is the fourth place where the soft shrinkage function appears naturally: As the maximum a posteriori estimator of a Laplace distributed random variable disturbed by white noise.

The assumption, that the wavelet coefficients are Laplace distributed may not be accurate. One can model the distribution of wavelet coefficients by

$$p(x) = Ce^{-\phi(x)}$$

where ϕ is a weight function. This distribution clearly leads to the maximum a posteriori estimator

$$\hat{x}(y) = \operatorname*{argmin}_x (y-x)^2 + 2C\sigma^2\phi(x)$$

which we already know from Section 4.1. In [ML99] different "heavy-tailed" distributions, namely generalized Gaussian distributions, of the wavelet coefficients are analyzed and the corresponding shrinkage functions are calculated. A generalized Gaussian distribution has the form

$$p(x) = \frac{a}{\sigma_p} e^{-\left(\frac{b|x|}{\sigma_p}\right)^p}$$

where p is the so called "shape parameter" of the distribution, a and b are constants and σ_p is the variance. The maximum a posteriori estimator under such a model for the wavelet coefficients is

$$\hat{x}(y) = \operatorname*{argmin}_x (y-x)^2 + 2b^p \frac{\sigma^2}{\sigma_p^p} |x|^p .$$

From Section 4.1 we know how this estimator looks like for different cases of p.

The equivalence of Bayesian denoising and variational denoising is formulated in the following theorem.

Theorem 5.2 (Equivalence of Bayesian and variational denoising). *Let $g \in L^2(I)$ be a noisy signal or image and let ψ be an orthogonal wavelet. Then the following denoising methods do the same.*

- *Suppose $g = f + \varepsilon$, where ε is white noise with variance σ and the wavelet coefficients of f with respect to ψ are generalized Gaussian distributed with shape parameter p, parameter b and variance σ_p. Estimate the wavelet coefficients of f by the maximum a posteriori estimator under the above assumptions.*

- *The function f is the minimizer of the following functional*

$$F(u) = \|u - g\|^2_{L^2(I)} + 2b^p \frac{\sigma^2}{\sigma_p^p} |u|^p_{B^{(1/p-1/2)d}_{p,p}(I)}$$

where the Besov seminorm is characterized through the wavelet coefficients with respect to ψ.

In some applications one may not be interested in removing white noise from signals but other disturbances. Thus one may be interested in other models for the distributions for the signal and the noise. We end this chapter with the following proposition which shows what happens if one switches the distributions of the signal and the noise.

Proposition 5.3 (One man's signal is the other man's noise). *Let f be the desired signal which is additively disturbed by noise ε: $g = f + \varepsilon$. Let the wavelet coefficients of f and the noise ε be distributed as $e^{-\phi(\cdot)}$ and $e^{-\psi(\cdot)}$ respectively with $\psi, \phi : \mathbb{R} \to \overline{\mathbb{R}}$ symmetric.*

Let the maximum a posterior estimator for the wavelet coefficients of f be

$$\hat{f}_\gamma(g_\gamma) = S(g_\gamma).$$

If one assumes that the distributions of the coefficients of f and ε are the other way round one get the maximum a posteriori estimator

$$\hat{f}_\gamma(g_\gamma) = (\mathrm{Id} - S)(g_\gamma).$$

The proof of this proposition follows from this simple lemma.

Lemma 5.4. *Let $\psi, \phi : \mathbb{R} \to \overline{\mathbb{R}}$ be symmetric and*

$$S(x) := \operatorname*{argmin}_w \{\psi(w - x) + \phi(w)\}, \quad R(x) := \operatorname*{argmin}_w \{\phi(w - x) + \psi(w)\}.$$

Then

$$S(x) + R(x) = x.$$

Proof. The proof is just a simple calculation:

$$\begin{aligned} x - R(x) &= x - \operatorname*{argmin}_w \{\psi(w) + \phi(w - x)\} \\ &= x + \operatorname*{argmin}_w \{\psi(-w) + \phi(-w - x)\} \\ &= x + \operatorname*{argmin}_w \{\psi(w) + \phi(w + x)\} \\ &= \operatorname*{argmin}_w \{\psi(w - x) + \phi(w)\} \\ &= S(x). \end{aligned}$$

□

CHAPTER 6

Summary

This thesis is a contribution to the investigation of equivalences of methods in image and signal processing. The method of soft wavelet shrinkage is in the center of the investigation and the thesis presents a collection of equivalence results for this method. The different equivalence results for soft wavelet shrinkage are not new but they have been published in different fields of mathematical image processing. This thesis provides an extensive study of the equivalence results and presents them in a unified framework.

The different equivalence results for soft wavelet shrinkage are collected in this theorem.

Theorem 6.1 (Equivalences to soft wavelet shrinkage). *Let $f : I \to \mathbb{R}$ be a signal or image and $g = f + \varepsilon$ where ε is white noise of variance σ. Further, let $\lambda > 0$ and $\{\psi_\gamma\}$ be an orthonormal wavelet base of $L^2(I)$. Then the following denoising methods are equivalent.*

- *$\bar{f}$ is the soft wavelet shrinkage of g, i. e.*

$$\bar{f} = \langle g | 1 \rangle + \sum_{\gamma \in \Gamma} S_\lambda(g_\gamma)\psi_\gamma.$$

- *$\bar{f}$ is the solution of the descent equation*

$$\partial_t f + \partial |f|_{B^{d/2}_{1,1}(I)} \ni 0, \;\; f(0) = g$$

  at time $t = \lambda$. (compare Section 3.2).

- *$\bar{f}$ is the minimizer of the functional*

$$\|g - f\|^2_{L^2(I)} + 2\lambda \, |f|_{B^{d/2}_{1,1}(I)}$$

(compare Section 4.1 or Section 4.2).

- *$\bar{f}$ is is given by "$g -$ Projection of g onto a $B^{-1}_{\infty,\infty}(I)$-ball of radius λ", i. e.*

$$\bar{f} = (\mathrm{Id} - \Pi_{\lambda C})(g)$$

with $C := \left\{ h \in L^2(I) \,\middle|\, |h|_{B^{-1}_{\infty,\infty}(I)} \leq 1 \right\}$. In other words: $\bar{f}$ is obtained by subtracting a texture component of size λ measured in $B^{-1}_{\infty,\infty}(I)$ (compare Section 4.2).

- *$\bar{f}$ is obtained by the maximum a posteriori estimator of the wavelet coefficients of f under the assumption, that the wavelet coefficients of f are distributed Laplacian with parameter $\eta = \frac{\lambda}{\sqrt{2\sigma^2}}$ (compare Chapter 5).*

Beside the topic of soft wavelet shrinkage, every of the equivalent descriptions leads to different generalizations. The approach which relates wavelet shrinkage to descent equations is easily extended to other shrinkage methods which involve isometrical and invertible transformations as it is shown for the Fourier transform in Chapter 3.

The description of wavelet shrinkage as a variational problem is probably the oldest one and dates back to 1992 [DL92a]. In this thesis a broad class of variational problems is treated. The variational functionals

$$\|u - f\|^2_{L^2(I)} + 2\lambda \, |u|^p_{B^s_{p,p}(I)}$$

are considered for all cases of s and p. The exact minimizers (with respect to a chosen wavelet basis in which the Besov semi norm is described) are calculated and with the help of methods from convex analysis it is shown, that they are in fact equivalent to wavelet shrinkage with different shrinkage functions. Furthermore, soft and hard wavelet shrinkage appeared as the special cases $p = 1$ resp. $p = 0$. A surprising result is the resulting interpolation between soft and hard shrinkage. The interpolating shrinkage functions are discontinuous for $p < 1$ and the behavior of the effective threshold is rather complicated (compare Section 4.1).

The almost similar variational functional

$$\|u - f\|^2_{L^2(I)} + 2\lambda \, |u|_{B^s_{p,p}(I)}$$

is treated as well in this thesis for all cases of s and $p \geq 1$ in Section 4.2. Again the minimizers are calculated, but with different methods. The interesting thing about this functional is not the explicit formula for the minimizer but the description in terms of projections. This leads to a new interpretation of shrinkage methods even for methods which do not involve wavelets or the basis expansions. It is shown, that for every functional as above where the penalty term is positively homogeneous the minimizer is of the form $u = (\mathrm{Id} - \Pi_{\lambda C})(f)$. In other words: To minimize such a variational functional one has to subtract a "part of size λ in the set C" of the given image or signal f. The set C is a kind of unit ball in the dual norm of the penalty term. This can also be applied for example for the total variation methods where the set C has a complicated structure (compare [AA03, Cha04, MH04]).

The last description of wavelet shrinkage which is treated in this thesis involves terms of Bayesian statistics and is described shortly in Chapter 5. It is shown, how certain assumptions on the noise model and the distribution of the wavelet coefficients lead to shrinkage methods. The relation to the variational methods as in Section 4.1 then is simple and obvious.

The field of equivalences in image and signal processing is still growing and there are many more problems to solve. This thesis presents an extensive study for different denoising methods and shows, that many of them are leading to the idea of shrinkage in a general sense.

BIBLIOGRAPHY

[AA03] Gilles Aubert and Jean-François Aujol. Modeling very oscillating signals. Application to image processing. Technical Report 4878, Institut National de Recherche en Informatique et en Automatique, July 2003.

[AC04] Jean-François Aujol and Antonin Chambolle. Dual norms and image decomposition models. Technical Report 5130, Institut National de Recherche en Informatique et en Automatique, March 2004.

[ALM92] Luis Alvarez, Pierre-Louis Lions, and Jean-Michel Morel. Image selective smoothing and edge detection by nonlinear diffusion II. *SIAM Journal on Numerical Analysis*, 29(3):845–866, 1992.

[BLM03] Kristian Bredies, Dirk A. Lorenz, and Peter Maass. Mathematical concepts of multiscale smoothing. University of Bremen, Preprint Series of the DFG-Schwerpunktprogramm 1114, April 2003.

[Bre73] H. Brezis. *Opérateurs maximaux monotones et semi-groups de contractions dans les espaces de Hilbert*, volume 5 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., 1973.

[BWWS03] Thomas Brox, Martin Welk, Joachim Weickert, and Gabriele Steidl. Equivalence results for TV diffusion and TV regularisation. In L. D. Griffin and M. Lillholm, editors, *Scale Space*

Methods in Computer Vision, number 2695 in Lecture Notes in Computer Science, pages 86–100. Springer, 2003.

[CD95] Ronald R. Coifman and David L. Donoho. Translation-invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Springer Lecture Notes in Statistics*, pages 125–150. Springer-Verlag, 1995.

[CDDD03] Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, and Ronald A. DeVore. Harmonic analysis of the space BV. *Revista Matematica Iberoamericana*, 19(1):235–263, 2003.

[CDLL98] Antonin Chambolle, Ronald A. DeVore, Namyong Lee, and Bradley J. Lucier. Nonlinear wavelet image processing: Variational problems, compression and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7:319–335, 1998.

[Cha04] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004.

[CL97] Antonin Chambolle and Pierre-Louis Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76:167–188, 1997.

[CL01] Antonin Chambolle and Bradley J. Lucier. Interpreting translation-invariant wavelet shrinkage as a new image smoothing scale space. *IEEE Transactions on Image Processing*, 10:993–1000, 2001.

[CLMC92] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.

[Coh03] Albert Cohen. *Numerical Analysis of Wavelet Methods*. Elsevier Science B. V., 2003.

[cS02] Levent Şendur and Ivan W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Transactions on Signal Processing*, 50(11):2744–2756, 2002.

[Dau92] Ingrid Daubechies. Ten Lectures on Wavelets. In *Cbms-Nsf Regional Conference Series in Applied Mathematics*, volume 61, 1992.

[DeV98] Ronald A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[DJ94] David L. Donoho and I.M. Johnstone. Ideal spatial adaption via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[DJ98] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

[DJKP96] David L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Annals of Statistics*, 24(2), 1996.

[DKW98] Ronald A. DeVore, G. C. Kyriazis, and P. Wang. Multiscale characterizations of Besov spaces on bounded domains. *Journal of Approximation Theory*, 2(93):273–292, 1998.

[DL92a] Ronald A. DeVore and Bradley J. Lucier. Fast wavelet techniques for near-optimal image processing. In *Communications - Fusing Command, Control and Intelligence*, volume 3 of *Military Communications Conference, MILCOM '92*, pages 129–135. IEEE, 1992.

[DL92b] Ronald A. DeVore and Bradley J. Lucier. Wavelets. *Acta Numerica*, pages 1–56, 1992.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*. Springer, 1993.

[Don95] David L. Donoho. Denoising via soft thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[DV97] David C. Dobson and Curtis R. Vogel. Convergence of an iterative method for total variation denoising. *SIAM Journal on Numerical Analysis*, 34:1779–1971, 1997.

[ET76] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. North-Holland, Amsterdam, 1976.

[FJW91] Michael Frazier, Björn Jawerth, and Guido Weiss. *Littlewood-Paley theory and the study of function spaces*. Number 79 in Regional Conference Series in Mathematics. American Mathematical Society, 1991.

[FN01] Mário A. T. Figueiredo and Robert D. Nowak. Wavelet-based image estimation: An empirical Bayes approach using Jefferey's noninformative prior. *IEEE Transactions on Image Processing*, 10(9):1322–1331, 2001.

[HK04] M. Hintermüller and K. Kunisch. Total bounded variation regularization as a bilaterally constrained optimization problem. *SIAM Journal on Applied Mathematics*, 64(4):1311–1333, 2004.

[Kee02] Stephen L. Keeling. Nonlinear anisotropic diffusion filtering for multiscale edge enhancement. *Inverse Problems*, 18(1):175–190, 2002.

[Kee03] Stephen L. Keeling. Total variation based convex filters for medical imaging. *Applied Mathematics and Computation*, 139:101–119, 2003.

[Kic97] Satyanad Kichenassamy. The Perona-Malik paradox. *SIAM Journal of Applied Mathematics*, 57:1328–1342, 1997.

[KK98] Bernd Kawohl and Nikolai Kutev. Maximum and comparison principle for one-dimensional anisotropic diffusion. *Mathematische Annalen*, 311:107–123, 1998.

[LMR97] Alfred Karl Louis, Peter Maass, and Andreas Rieder. *Wavelets: Theory and Application*. Wiley, Chichester, 1997.

[Lor02] Dirk A. Lorenz. Methoden der Multiskalenglättung. Master's thesis, Universität Bremen, 2002.

[Lor04a] Dirk A. Lorenz. Solving variational methods in image processing via projections - a common view on TV-denoising and wavelet shrinkage. University of Bremen, Preprint Series of the DFG-Schwerpunktprogramm 1114, 2004.

[Lor04b] Dirk A. Lorenz. Variational denoising in Besov spaces and interpolation of hard and soft wavelet shrinkage. University of Bremen, Preprint Series of the DFG-Schwerpunktprogramm 1114, 2004.

[Mal99] Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, USA, 1999.

[Mey92] Yves Meyer. *Wavelets and Operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1992.

[Mey01] Yves Meyer. *Oscillating patterns in image processing and nonlinear evolution equations*, volume 22 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2001. The fifteenth Dean Jacqueline B. Lewis memorial lectures.

[MH04] Yves Meyer and Ali Haddad. Variational methods in image processing. Extended notes from *Wavelet and Multifractal Analysis* Cargese, 2004.

[ML99] Pierre Moulin and Juan Liu. Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Transactions on Information Theory*, 45(3):909–919, 1999.

[MW03] Pavel Mrázek and Joachim Weickert. Rotationally invariant wavelet shrinkage. In B. Michaelis and G. Krell, editors, *Pattern Recognition. 25th DAGM Symposium*, volume 2781 of *LNCS*, pages 156–163. Springer, 2003.

[MWS03] Pavel Mrázek, Joachim Weickert, and Gabriele Steidl. Correspondences between wavelet shrinkage and nonlinear diffusion. In L. D. Griffin and M. Lillholm, editors, *Scale Space Methods in Computer Vision*, number 2695 in Lecture Notes in Computer Science, pages 101–116. Springer, 2003.

[OOS04] Andreas Obereder, Stanley Osher, and Otmar Scherzer. On the use of dual norms in bounded variation type regularization. Preprint, Infmath Imaging, University Innsbruck, 2004.

[OSV03] Stanley J. Osher, Andrés Solé, and Luminita A. Vese. Image decomposition and restoration using total variation minimization and the H^{-1} norm. *Multiscale Modelling and Simulation*, 1(3):349–370, 2003.

[PM90] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[ROF92] Leonid I. Rudin, Stanley J. Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[RS96] Thomas Runst and Winfried Sickel. *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*. de Gruyter Series in Nonlinear Analysis and Applications. Walter de Gruyter, 1996.

[Rud87] Walter Rudin. *Real and Complex Analyis*. McGraw-Hill, 3rd edition, 1987.

[RW98] R. Tyrell Rockafellar and Roger J-B. Wets. *Variational Analysis*. Springer, 1998.

[SA96] Eero P. Simoncelli and Edward H. Adelson. Noise removal via Bayesian wavelet coring. In *3rd IEEE International Conference on Image Processing*, volume 1, pages 379–382. IEEE Signal Processing Society, September 1996.

[SC96] David M. Strong and Tony F. Chan. Exact solutions to total variation regularization problems. UCLA Math Department CAM Report 96-41, October 1996.

[SC03] David M. Strong and Tony F. Chan. Edge-preserving and scale dependent properties of total variation regularization. *Inverse Problems*, 19:165–187, 2003.

[Tri92] Hans Triebel. *Theory of Function Spaces II*. Monographs in Mathematics. Birkhäuser, 1992.

[Ves01] Luminita A. Vese. A study in the BV space of a denoising-deblurring variational problem. *Applied Mathematics and Optimization*, 44:131–161, 2001.

[Wil00] Elke Wilczok. New uncertainty principles for the continuous Gabor transform and the continuous wavelet transform. *Documenta Mathematica*, 5:201–226, 2000.

[WXHC91] J. B. Weaver, Yansun Xu, D. M. Healy, and L. D. Cromwell. Filtering noise from images with wavelet transforms. *Magnetic Resonance in Medicine*, 21:288–295, 1991.